A METHOD FOR IDENTIFYING A SUBSET OF COMPONENTS OF A SYSTEM

FIELD OF THE INVENTION

The present invention relates to a method and apparatus for identifying components of a system from data generated from samples from the system, which components are capable of predicting a feature of the sample within the system and, particularly, but not exclusively, the present invention relates to a method and apparatus for identifying components of a biological system from data generated by a biological method, which components are capable of predicting a feature of interest associated with a sample applied to the biological system.

BACKGROUND OF THE INVENTION

There are any number of systems in existence that can be classified according to one or more features thereof. The term "system" as used throughout this specification is considered to include all types of systems from which data (e.g. statistical data) can be obtained. Examples of such systems include chemical systems, financial systems and geological systems. It is desirable to be able to utilise data obtained from the systems to identify particular features of samples from the system; for instance, to assist with analysis of financial system to identify groups such as those who have good credit and those who are a credit risk. Often the data obtained from the systems is relatively large and therefore it is desirable to identify components of the systems from the data, the components being predictive of the particular features of the samples from the system. However, when the data is relatively large it can be difficult to identify the components because there is a large amount of data to process, the majority of which may not provide any indication or little indication of the features of a particular sample from which the data is

taken. Furthermore, components that are identified using a training sample are often ineffective at identifying features on test sample data when the test sample data has a high degree of variability relative to the training sample data. This is often the case in situations when, for example, data is obtained from many different sources, as it is often difficult to control the conditions under which the data is collected from each individual source.

An example of a type of system where these problems are particularly pertinent, is a biological system, in which the components could include, for example, particular genes or proteins. Recent advances in biotechnology have resulted in the development of biological methods for large scale screening of systems and analysis of samples. Such methods include, for example, microarray analysis using DNA or RNA, proteomics analysis, proteomics electrophoresis gel analysis, and high throughput screening techniques. These types of methods often result in the generation of data that can have up to 30,000 or more components for each sample that is tested.

It is highly desirable to be able identify features of interest in samples from biological systems. For example, to classify groups such as "diseased" and "non-diseased". Many of these biological methods would be useful as diagnostic tools predicting features of a sample in the biological systems. For example, identifying diseases by screening tissues or body fluids, or as tools for determining, for example, the efficacy of pharmaceutical compounds.

Use of biological methods such as biotechnology arrays in such applications to date has been limited due to the large amount of data that is generated from these types of methods, and the lack of efficient methods for screening the data for meaningful results. Consequently, analysis of

biological data using existing methods is time consuming, prone to false results and requires large amounts of computer memory if a meaningful result is to be obtained from the data. This is problematic in large scale screening scenarios where rapid and accurate screening is required.

It is therefore desirable to have a method, in particular for analysis of biological data, and more generally, for an improved method of analysing data from a system in order to predict a feature of interest for a sample from the system.

SUMMARY OF THE INVENTION

According to a first aspect of the present invention, there is provided a method of identifying a subset of components of a system based on data obtained from the system using at least one training sample from the system, the method comprising the steps of:

obtaining a linear combination of components of the system and weightings of the linear combination of components, the weightings having values based on the data obtained from the system using the at least one training sample, the at least one training sample having a known feature;

obtaining a model of a probability distribution of the known feature, wherein the model is conditional on the linear combination of components;

obtaining a prior distribution for the weighting of the linear combination of the components, the prior distribution comprising a hyperprior having a high probability density close to zero, the hyperprior being such that it is not a Jeffreys hyperprior;

combining the prior distribution and the model to generate a posterior distribution; and

identifying the subset of components based on a set of the weightings that maximise the posterior distribution.

The method utilises training samples having the known feature in order to identify the subset of components which can predict a feature for a training sample. Subsequently, knowledge of the subset of components can be used for tests, for example clinical tests, to predict a feature such as whether a tissue sample is malignant or benign, or what is the weight of a tumour, or provide an estimated time for survival of a patient having a particular condition.

The term "feature" as used throughout this specification refers to any response or identifiable trait or character that is associated with a sample. For example, a feature may be a particular time to an event for a particular sample, or the size or quantity of a sample, or the class or group into which a sample can be classified.

Preferably, the step of obtaining the linear combination comprises the step of using a Bayesian statistical method to estimate the weightings.

Preferably, the method further comprises the step of making an apriori assumption that a majority of the components are unlikely to be components that will form part of the subset of components.

The apriori assumption has particular application when there are a large amount of components obtained from the system. The apriori assumption is essentially that the majority of the weightings are likely to be zero. The model is constructed such that with the apriori assumption in mind, the weightings are such that the posterior probability of the weightings given the observed data is maximised. Components having a weighting below a pre-determined threshold (which will be the majority of them in accordance with the apriori assumption) are ignored. The process is iterated until the correct diagnostic components are identified. Thus, the method has the potential to be quick,

mainly because of the apriori assumption, which results in rapid elimination of the majority of components.

Preferably, the hyperprior comprises one or more adjustable parameter that enable the prior distribution near zero to be varied.

Most features of a system typically exhibit a probability distribution, and the probability distribution of a feature can be modelled using statistical models that are based on the data generated from the training samples. The present invention utilises statistical models that model the probability distribution for a feature of interest or a series of features of interest. Thus, for a feature of interest having a particular probability distribution, an appropriate model is defined that models that distribution.

Preferably, the method comprise a mathematical equation in the form of a likelihood function that provides the probability distribution based on data obtained from the at least one training sample.

Preferably, the likelihood function is based on a previously described model for describing some probability distribution.

Preferably, the step of obtaining the model comprises the step of selecting the model from a group comprising a multinomial or binomial logistic regression, generalised linear model, Cox's proportional hazards model, accelerated failure model and parametric survival model.

In a first embodiment, the likelihood function is based on the multinomial or binomial logistic regression. The binomial or multinomial logistic regression preferably models a feature having a multinomial or binomial distribution. A binomial distribution is a statistical

distribution having two possible classes or groups such as an on/off state. Examples of such groups include dead/alive, improved/not improved, depressed/not depressed. A multinomial distribution is a generalisation of the binomial distribution in which a plurality of classes or groups are possible for each of a plurality of samples, or in other words, a sample may be classified into one of a plurality of classes or groups. Thus, by defining a likelihood function based on a multinomial or binomial logistic regression, it is possible to identify subsets of components that are capable of classifying a sample into one of a plurality of pre-defined groups or classes. To do this, training samples are grouped into a plurality of sample groups (or "classes") based on a predetermined feature of the training samples in which the members of each sample group have a common feature and are assigned a common group identifier. A likelihood function is formulated based on a multinomial or binomial logistic regression conditional on the linear combination (which incorporates the data generated from the grouped training samples). The feature may be any desired classification by which the training samples are to be grouped. For example, the features for classifying tissue samples may be that the tissue is normal, malignant, benign, a leukemia cell, a healthy cell, that the training samples are obtained from the blood of patients having or not having a certain condition, or that the training samples are from a cell from one of several types of cancer as compared to a normal cell.

In the first embodiment, the likelihood function based on the multinomial or binomial logistic regression is of the form:

$$L = \prod_{i=1}^{n} \left(\prod_{g=1}^{G-1} \left\{ \frac{e^{x_i^T \beta_g}}{\left(1 + \sum_{g=1}^{G-1} e^{x_i^T \beta_g} \right)} \right\}^{e_{ik}} \left\{ \frac{1}{1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}} \right\}^{e_{iG}} \right)$$

wherein

$x_i^T \beta_g$ is a linear combination generated from input data from training sample i with component weights β_g;

x_i^T is the components for the i^{th} Row of X and β_g is a set of component weights for sample class g; and

X is data from n training samples comprising p components and the e_{ik} are defined further in this specification.

In a second embodiment, the likelihood function is based on the ordered categorical logistic regression. The ordered categorical logistic regression models a binomial or multinomial distribution in which the classes are in a particular order (ordered classes such as for example, classes of increasing or decreasing disease severity). By defining a likelihood function based on an ordered categorical logistic regression, it is possible to identify a subset of components that is capable of classifying a sample into a class wherein the class is one of a plurality of predefined ordered classes. By defining a series of group indentifiers in which each group identifier corresponds to a member of an ordered class, and grouping the training samples into one of the ordered classes based on predetermined features of the training samples, a likelihood function can be formulated based on a categorical ordered logistic regression which is conditional on the linear combination (which incorporates the data generated from the grouped training samples).

In the second embodiment, the likelihood function based on the categorical ordered logistic regression is of the form:

$$l = \prod_{i=1}^{N} \prod_{k=1}^{G-1} \left(\frac{\gamma_{ik}}{\gamma_{ik+1}} \right)^{r_{ik}} \left(\frac{\gamma_{ik+1} - \gamma_{ik}}{\gamma_{ik+1}} \right)^{r_{ik+1} - r_{ik}}$$

Wherein

γ_{ik} is the probability that training sample i belongs to a class with identifier less than or equal to k (where the

total of ordered classes is G).The r_i is defined further in the document.

In a third embodiment of the present invention, the likelihood function is based on the generalised linear model. The generalised linear model preferably models a feature that is distributed as a regular exponential family of distributions. Examples of regular exponential family of distributions include normal distribution, guassian distribution, poisson distribution, gamma distribution and inverse gaussian distribution. Thus, in another embodiment of the method of the invention, a subset of components is identified that is capable of predicting a predefined characteristic of a sample which has a distribution belonging to a regular exponential family of distributions. In particular by defining a generalised linear model which models the characteristic to be predicted. Examples of a characteristic that may be predicted using a generalised linear model include any quantity of a sample that exhibits the specified distribution such as, for example, the weight, size or other dimensions or quantities of a sample.

In the third embodiment, the generalised linear model is of the form:

$$L = \log p(y \mid \beta, \phi) = \sum_{i=1}^{N} \{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \}$$

where $y = (y_1, \ldots, y_n)^T$ and $a_i(\phi) = \phi / w_i$ with the w_i being a fixed set of known weights and ϕ a single scale parameter. The other terms in this expression are defined later in this document.

In a fourth embodiment, the method of the present invention may be used to predict the time to an event for a sample by utilising the likelihood function that is based on a hazard model, which preferably estimates the probability of a time to an event given that the event has not taken place at the

time of obtaining the data. In the fourth embodiment, the likelihood function is selected from the group comprising a Cox's proportional hazards model, parametric survival model and accelerated failure times model. Cox's proportional hazards model permits the time to an event to be modelled on a set of components and component weights without making restrictive assumptions about time. The accelerated failure model is a general model for data consisting of survival times in which the component measurements are assumed to act multiplicatively on the time-scale, and so affect the rate at which an individual proceeds along the time axis. Thus, the accelerated survival model can be interpreted in terms of the speed of progression of, for example, disease. The parametric survival model is one in which the distribution function for the time to an event (eg survival time) is modelled by a known distribution or has a specified parametric formulation. Among the commonly used survival distributions are the Weibull, exponential and extreme value distributions.

In the fourth embodiment, a subset of components capable of predicting the time to an event for a sample is identified by defining a likelihood based on Cox's proportional standards model, a parametric survival model or an accelerated survival times model, which comprises measuring the time elapsed for a plurality of samples from the time the sample is obtained to the time of the event.

In the fourth embodiment, the likelihood function for predicting the time to an event is of the form:

$$Log\ (Partial)\ Likelihood\ = \sum_{i=1}^{N} g_i\left(\underset{\sim}{\beta}, \underset{\sim}{\varphi}; X, \underset{\sim}{y}, \underset{\sim}{c}\right)$$

where $\underset{\sim}{\beta}' = (\beta_1, \beta_2, \cdots, \beta_p)$ and $\underset{\sim}{\varphi}' = (\varphi_1, \varphi_2, \cdots, \varphi_q)$ are the model parameters, y is a vector of observed times and c is an indicator vector which indicates whether a time is a true

survival time or a censored survival time.

In the fourth embodiment, the likelihood function based on Cox's proportional hazards model is of the form:

$$l\left(\underset{\sim}{t} \mid \underset{\sim}{\beta}\right) = \prod_{j=1}^{N} \left(\frac{exp\left(Z_j \underset{\sim}{\beta}\right)}{\sum_{i \in \Re_j} exp\left(Z_i \underset{\sim}{\beta}\right)} \right)^{d_j}$$

where the observed times are be ordered in increasing magnitude denoted as $\underset{\sim}{t} = \left(t_{(1)}, t_{(2)}, \cdots t_{(N)}\right)$, $t_{(i+1)} > t_{(i)}$. and Z denotes the Nxp matrix that is the re-arrangement of the rows of X where the ordering of the rows of Z corresponds to the ordering induced by the ordering of $\underset{\sim}{t}$. Also $\underset{\sim}{\beta}^T = (\beta_1, \beta_2, \cdots, \beta_n)$, Z_j = the j^{th} row of Z, and $\Re_j = \{i : i = j, j+1, \cdots, N\}$ = the risk set at the j^{th} ordered event time $t_{(j)}$.

In the fourth embodiment, wherein the likelihood function is based on the Parametric Survival model it is of the form:

$$L = \sum_{i=1}^{N} \left\{ c_i log\left(\mu_i\right) - \mu_i + c_i \left(log\left(\frac{\lambda\left(y_i\right)}{\Lambda\left(y_i ; \underset{\sim}{\varphi}\right)} \right) \right) \right\}$$

where $\mu_i = \Lambda\left(y_i ; \underset{\sim}{\varphi}\right) exp\left(X_i \underset{\sim}{\beta}\right)$ and Λ denotes the integrated parametric hazard function.

For any defined models, the weightings are typically estimated using a Bayesian statistical model (Kotz and Johnson, 1983) in which a posterior distribution of the component weights is formulated which combines the likelihood function and a prior distribution. The component weightings are estimated by maximising the posterior

distribution of the weightings given the data generated for the at least one training sample. Thus, the objective function to be maximised consists of the likelihood function based on a model for the feature as discussed above and a prior distribution for the weightings.

Preferably, the prior distribution is of the form:

$$p(\beta) = \int_{v^2} p(\beta \mid v^2)\, p(v^2)\, dv^2$$

wherein v is a p x 1 vector of hyperparameters, and where $p(\beta \mid v^2)$ is $N(0, \mathrm{diag}\{v^2\})$ and $p(v^2)$ is some hyperprior distribution for v^2.

Preferably, the hyperprior comprises a gamma distribution with a specified shape and scale parameter.

This hyperprior distribution (which is preferably the same for all embodiments of the method) may be expressed using different notational conventions, and in the detailed description of the embodiments (see below), the following notational conventions are adopted merely for convenience for the particular embodiment:

As used herein, when the likelihood function for the probability distribution is based on a multinomial or binomial logistic regression, the notation for the prior distribution is:

$$P(\beta_1, \ldots, \beta_{i-1}) = \int_{\tau^2} \prod_{g=1}^{G-1} P(\beta_g \mid \tau_g^2)\, P(\tau_g^2)\, d\tau^2$$

where $\beta^T = (\beta_1^T, \ldots \beta_{G-1}^T)$ and $\tau^T = (\tau_1^T, \ldots, \tau_{G-1}^T)$.

and $p(\beta_g|\tau_g^2)$ is $N(0,\text{diag}\{\tau_g^2\})$ and $P(\tau_g^2)$ is some hyperprior distribution for τ_g^2.

As used herein, when the likelihood function for the probability distribution is based on a categorical ordered logistic regression, the notation for the prior distribution is:

$$P(\beta_1,\beta_2,\cdots,\beta_n)=\int_{\tau}\prod_{i=1}^{N}P(\beta_i|\nu_i^2)P(\nu_i^2)d\nu^2$$

where $\beta_1,\beta_2,\cdots,\beta_n$ are component weights, $P(\beta_i|\nu_i)$ is $N(0,\nu_i')$ and $P(\nu_i)$ some hyperprior distribution for ν_i.

As used herein, when the likelihood function for the distribution is based on a generalised linear model, the notation for the prior distribution is:

$$p(\beta)=\int_{\tau^2}p(\beta\,|\nu^2)p(\nu^2)d\nu^2$$

wherein ν is a p x 1 vector of hyperparameters, and where $p(\beta\,|\nu^2)$ is $N(0,\text{diag}\{\nu^2\})$ and $p(\nu^2)$ is some prior distribution for ν^2.

As used herein, when the likelihood function for the distribution is based on a hazard model, the notation for the prior distribution is:

where $p(\beta^*|\tau)$ is $N(0,\text{diag}\{\tau^2\})$ and $p(\tau)$ some hyperprior distribution for τ.

The prior distribution comprises a hyperprior that ensures that zero weightings are used whenever possible.

$$p(\beta^*) = \int_{\tau} p(\beta^* | \tau\ ) p(\tau\ ) d\tau$$

In an alternative embodiment, the hyperprior is an inverse gamma distribution in which each $t_i^2 = 1/v_i^2$ has an independent gamma distribution.

In a further alternative embodiment, the hyperprior is a gamma distribution in which each v_i^2, τ_i or τ_i^2 (depending on the context) has an independent gamma distribution.

As discussed previously , the prior distribution and the likelihood function are combined to generate a posterior distribution. The posterior distribution is preferably of the form:

$$p(\beta\, \varphi\, v | y) \alpha\ L(y | \beta\, \varphi) p(\beta | v) p(v)$$

or

$$P(\underset{\sim}{\beta}, \underset{\sim}{\varphi}, \underset{\sim}{\tau} | \underset{\sim}{y}) \alpha L(\underset{\sim}{y} | \underset{\sim}{\beta}, \underset{\sim}{\varphi}) P(\underset{\sim}{\beta} | \underset{\sim}{\tau}) P(\underset{\sim}{\tau})$$

wherein $L(\underset{\sim}{y} | \underset{\sim}{\beta}, \underset{\sim}{\varphi})$ is the likelihood function.

Preferably, the step of identifying the subset of components comprises the step of using an iterative procedure such that the probability density of the posterior distribution is maximised.

During the iterative procedure, component weightings having a value less than a pre-determined threshold are eliminated, preferably by setting those component weights to zero. This results in the substantially elimination of the

corresponding component.

Preferably, the iterative procedure is an EM algorithm.

The EM algorithm produces a sequence of component weighting estimates that converge to give component the weightings that maximise the probability density of the posterior distribution. The EM algorithm consists of two steps, known as the E or Expectation step and the M, or Maximisation step. In the E step, the expected value of the log-posterior function conditional on the observed data is determined. In the M step, the expected log-posterior function is maximised to give updated component weight estimates that increase the posterior. The two steps are alternated until convergence of the E step and the M step is achieved, or in other words, until the expected value and the maximised value of the expected log-posterior function converges.

It is envisaged that the method of the present invention may be applied to any system from which measurements can be obtained, and preferably systems from which very large amounts of data are generated. Examples of systems to which the method of the present invention may be applied include biological systems, chemical systems, agricultural systems, weather systems, financial systems including, for example, credit risk assessment systems, insurance systems, marketing systems or company record systems, electronic systems, physical systems, astrophysics systems and mechanical systems. For example, in a financial system, the samples may be particular stock and the components may be measurements made on any number of factors which may affect stock prices such as company profits, employee numbers, rainfall values in various cities, number of shareholders etc.

The method of the present invention is particularly suitable for use in analysis of biological systems. The method of the present invention may be used to identify subsets of components for classifying samples from any biological system which produces measurable values for the components and in which the components can be uniquely labelled. In other words, the components are labelled or organised in a manner which allows data from one component to be distinguished from data from another component. For example, the components may be spatially organised in, for example, an array which allows data from each component to be distinguished from another by spatial position, or each component may have some unique identification associated with it such as an identification signal or tag. For example, the components may be bound to individual carriers, each carrier having a detectable identification signature such as quantum dots (see for example, Rosenthal, 2001, Nature Biotech 19: 621-622; Han et al. (2001) Nature Biotechnology 19: 631-635), fluorescent markers (see for example, Fu et al, (1999) Nature Biotechnology 17: 1109-1111), bar-coded tags (see for example, Lockhart and trulson (2001) Nature Biotechnology 19: 1122-1123).

In a particularly preferred embodiment, the biological system is a biotechnology array. Examples of biotechnology arrays include oligonucleotide arrays, DNA arrays, DNA microarrays, RNA arrays, RNA microarrays, DNA microchips, RNA microchips, protein arrays, protein microchips, antibody arrays, chemical arrays, carbohydrate arrays, proteomics arrays, lipid arrays. In another embodiment, the biological system may be selected from the group including, for example, DNA or RNA electrophoresis gels, protein or proteomics electrophoresis gels, biomolecular interaction analysis such as Biacore analysis, amino acid analysis, ADMETox screening (see for example High-throughput ADMETox estimation: In Vitro and In Silico approaches (2002), Ferenc Darvas and Gyorgy Dorman (Eds), Biotechniques Press),

protein electrophoresis gels and proteomics electrophoresis gels.

The components may be any measurable component of the system. In the case of a biological system, the components may be, for example, genes or portions thereof, DNA sequences, RNA sequences, peptides, proteins, carbohydrate molecules, lipids or mixtures thereof, physiological components, anatomical components, epidemiological components or chemical components.

The training samples may be any data obtained from a system in which the feature of the sample is known. For example, training samples may be data generated from a sample applied to a biological system. For example, when the biological system is a DNA microarray, the training sample may be data obtained from the array following hybridisation of the array with RNA extracted from cells having a known feature, or cDNA synthesised from the RNA extracted from cells, or if the biological system is a proteomics electrophoresis gel, the training sample may be generated from a protein or cell extract applied to the system.

It is envisaged that an embodiment of a method of the present invention may be used in re-evaluating or evaluating test data from subjects who have presented mixed results in response to a test treatment. Thus, there is a second aspect to the present invention.

The second aspect provides a method for identifying a subset of components of a subject which are capable of classifying the subject into one of a plurality of predefined groups, wherein each group is defined by a response to a test treatment, the method comprising the steps of:

exposing a plurality of subjects to the test treatment and grouping the subjects into response groups based on responses to the treatment;

measuring components of the subjects; and

identifying a subset of components that is capable of classifying the subjects into response groups using a statistical analysis method.

Preferably, the statistical analysis method comprises the method according to the first aspect of the present invention.

Once a subset of components has been identified, that subset can be used to classify subjects into groups such as those that are likely to respond to the test treatment and those that are not. In this manner, the method of the present invention permits treatments to be identified which may be effective for a fraction of the population, and permits identification of that fraction of the population that will be responsive to the test treatment.

According to a third aspect of the present invention, there is provided an apparatus for identifying a subset of components of a subject, the subset being capable of being used to classify the subject into one of a plurality of predefined response groups wherein each response group, is formed by exposing a plurality of subjects to a test treatment and grouping the subjects into response groups based on the response to the treatment, the apparatus comprising:

an input for receiving measured components of the subjects; and

processing means operable to identify a subset of components that is capable of being used to classify the subjects into response groups using a statistical analysis method.

Preferably, the statistical analysis method comprises the method according to the first or second aspect.

According to a fourth aspect of the present invention, there is provided a method for identifying a subset of components of a subject that is capable of classifying the subject as being responsive or non-responsive to treatment with a test compound, the method comprising the steps of:

exposing a plurality of subjects to the test compound and grouping the subjects into response groups based on each subjects response to the test compound;

measuring components of the subjects; and

identifying a subset of components that is capable of being used to classify the subjects into response groups using a statistical analysis method.

Preferably, the statistical analysis method comprises the method according to the first aspect.

According to a fifth aspect of the present invention, there is provided an apparatus for identifying a subset of components of a subject, the subset being capable of being used to classify the subject into one of a plurality of predefined response groups wherein each response group is formed by exposing a plurality of subjects to a compound and grouping the subjects into response groups based on the response to the compound, the apparatus comprising:

an input operable to receive measured components of the subjects;

processing means operable to identify a subset of components that is capable of classifying the subjects into response groups using a statistical analysis method.

Preferably, the statistical analysis method comprises the method according to the first or second aspect of the present invention.

The components that are measured in the second to fifth aspects of the invention may be, for example, genes or small nucleotide polymorphisms (SNPs), proteins, antibodies,

carbohydrates, lipids or any other measurable component of the subject.

In a particularly embodiment of the fifth aspect, the compound is a pharmaceutical compound or a composition comprising a pharmaceutical compound and a pharmaceutically acceptable carrier.

The identification method of the present invention may be implemented by appropriate computer software and hardware.

According to a sixth aspect of the present invention, there is provided an apparatus for identifying a subset of components of a system from data generated from the system from a plurality of samples from the system, the subset being capable of being used to predict a feature of a test sample, the apparatus comprising:

a processing means operable to:

obtain a linear combination of components of the system and obtain weightings of the linear combination of components, each of the weightings having a value based on data obtained from at least one training sample, the at least one training sample having a known feature;

obtaining a model of a probability distribution of a second feature, wherein the model is conditional on the linear combination of components;

obtaining a prior distribution for the weightings of the linear combination of the components, the prior distribution comprising an adjustable hyperprior which allows the prior probability mass close to zero to be varied wherein the hyperprior is not a Jeffrey's hyperprior;

combining the prior distribution and the model to generate a posterior distribution; and

identifying the subset of components having component weights that maximize the posterior distribution.

Preferably, the processing means comprises a computer arranged to execute software.

According to a seventh aspect of the present invention, there is provided a computer program which, when executed by a computing apparatus, allows the computing apparatus to carry out the method according to the first aspect of the present invention.

The computer program may implement any of the preferred algorithms and method steps of the first or second aspect of the present invention which are discussed above.

According to an eighth aspect of the present invention, there is provided a computer readable medium comprising the computer program according with the seventh aspect of the present invention.

According to a ninth aspect of the present invention, there is provided a method of testing a sample from a system to identify a feature of the sample, the method comprising the steps of testing for a subset of components that are diagnostic of the feature, the subset of components having been determined by using the method according to the first or second aspect of the present invention.

Preferably, the system is a biological system.

According to a tenth aspect of the present invention, there is provided an apparatus for testing a sample from a system to determine a feature of the sample, the apparatus comprising means for testing for components identified in accordance with the method of the first or second aspect of the present invention.

According to an eleventh aspect of the present invention, there is provided a computer program which, when executed by

on a computing device, allows the computing device to carry out a method of identifying components from a system that are capable of being used to predict a feature of a test sample from the system, and wherein a linear combination of components and component weights is generated from data generated from a plurality of training samples, each training sample having a known feature, and a posterior distribution is generated by combining a prior distribution for the component weights comprising an adjustable hyperprior which allows the probability mass close to zero to be varied wherein the hyperprior is not a Jeffrey's hyperprior, and a model that is conditional on the linear combination, to estimate component weights which maximise the posterior distribution.

Where aspects of the present invention are implemented by way of a computing device, it will be appreciated that any appropriate computer hardware e.g. a PC or a mainframe or a networked computing infrastructure, may be used.

According to a twelfth aspect of the present invention, there is provided a method of identifying a subset of components of a biological system, the subset being capable of predicting a feature of a test sample from the biological system, the method comprising the steps of:

obtaining a linear combination of components of the system and weightings of the linear combination of components, each of the weightings having a value based on data obtained from at least one training sample, the at least one training sample having a known first feature;

obtaining a model of a probability distribution of a second feature, wherein the model is conditional on the linear combination of components;

obtaining a prior distribution for the weightings of the linear combination of the components, the prior distribution comprising an adjustable hyperprior which allows the probability mass close to zero to be varied;

combining the prior distribution and the model to generate a posterior distribution; and

identifying the subset of components based on the weightings that maximize the posterior distribution.

BRIEF DESCRIPTION OF THE DRAWINGS

Notwithstanding any other embodiments that may fall within the scope of the present invention, an embodiment of the present invention will now be described, by way of example only, with reference to the accompanying figures, in which:

figure 1 provides a flow chart of a method according to the embodiment of the present invention;

figure 2 provides a flow chart of another method according to the embodiment of the present invention;

figure 3 provides a block diagram of an apparatus according to the embodiment of the present invention;

figure 4 provides a flow chart of a further method according to the embodiment of the present invention;

figure 5 provides a flow chart of an additional method according to the embodiment of the present invention; and

figure 6 provides a flow chart of yet another method according to the embodiment of the present invention.

DETAILED DESCRIPTION OF AN EMBODIMENT

The embodiment of the present invention identifies a relatively small number of components which can be used to identify whether a particular training sample has a feature. The components are "diagnostic" of that feature, or enable discrimination between samples having a different feature.

The number of components selected by the method can be controlled by the choice of parameters in the hyperprior. It is noted that the hyperprior is a gamma distribution with a specified shape and scale parameter. Essentially, from all the data which is generated from the system, the method of the present invention enables identification of a relatively small number of components which can be used to test for a particular feature. Once those components have been identified by this method, the components can be used in future to assess new samples. The method of the present invention utilises a statistical method to eliminate components that are not required to correctly predict the feature.

The inventors have found that component weightings of a linear combination of components of data generated from the training samples can be estimated in such a way as to eliminate the components that are not required to correctly predict the feature of the training sample. The result is that a subset of components are identified which can correctly predict the feature of the training sample. The method of the present invention thus permits identification from a large amount of data a relatively small and controllable number of components which are capable of correctly predicting a feature.

The method of the present invention also has the advantage that it requires usage of less computer memory than prior art methods. Accordingly, the method of the present invention can be performed rapidly on computers such as, for example, laptop machines. By using less memory, the method of the present invention also allows the method to be performed more quickly than other methods which use joint (rather than marginal) information on components for analysis of, for example, biological data.

The method of the present invention also has the advantage that it uses joint rather than marginal information on components for analysis.

A first embodiment relating to a multiclass logistic regression model will now be described.

A. Multi Class Logistic regression model

The method of this embodiment utilises the training samples in order to identify a subset of components which can classify the training samples into pre-defined groups. Subsequently, knowledge of the subset of components can be used for tests, for example clinical tests, to classify samples into groups such as disease classes. For example, a subset of components of a DNA microarray may be used to group clinical samples into clinically relevant classes such as, for example, healthy or diseased.

In this way, the present invention identifies preferably a small and controllable number of components which can be used to identify whether a particular training sample belongs to a particular group. The selected components are "diagnostic" of that group, or enable discrimination between groups. Essentially, from all the data which is generated from the system, the method of the present invention enables identification of a small number of components which can be used to test for a particular group. Once those components have been identified by this method, the components can be used in future to classify new samples into the groups. The method of the present invention preferably utilises a statistical method to eliminate components that are not required to correctly identify the group the sample belongs to.

The samples are grouped into sample groups (or "classes") based on a pre-determined classification. The classification

may be any desired classification by which the training samples are to be grouped. For example, the classification may be whether the training samples are from a leukemia cell or a healthy cell, or that the training samples are obtained from the blood of patients having or not having a certain condition, or that the training samples are from a cell from one of several types of cancer as compared to a normal cell.

In one embodiment, the input data is organised into an n x p data matrix $X=(x_{ij})$ with n training samples and p components. Typically, p will be much greater than n.

In another embodiment, data matrix X may be replaced by an n x n kernel matrix K to obtain smooth functions of X as predictors instead of linear predictors. An example of the kernel matrix K is $k_{ij}=\exp(-0.5*(x_i-x_j)^t(x_i-x_j)/\sigma^2)$ where the subscript on x refers to a row number in the matrix X. Ideally, subsets of the columns of K are selected which give sparse representations of these smooth functions.

Associated with each sample class (group) may be a class label y_i, where $y_i=k, k\in\{1,...,G\}$, which indicates which of G sample classes a training sample belongs to. We write the nx1 vector with elements y_i as $y_.$. Given the vector $\underset{\sim}{y}$ we can define indicator variables

$$e_{ig}=\begin{cases}1, & y_i=g\\ 0, & \text{otherwise}\end{cases} \qquad \text{(A1)}$$

In one embodiment, the component weights are estimated using a Bayesian statistical model (see Kotz and Johnson, 1983). Preferably, the weights are estimated by maximising the posterior distribution of the weights given the data generated from each training sample. This results in an objective function to be maximised consisting of two parts. The first part a likelihood function and the second a prior distribution for the weights which ensures that zero weights are preferred whenever possible. In a preferred embodiment,

the likelihood function is derived from a multiclass logistic model. Preferably, the likelihood function is computed from the probabilities:

$$p_{ig} = \frac{e^{x_i^T \beta_g}}{\left(1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}\right)}, g = 1, \ldots, G-1 \qquad \text{(A2)}$$

and

$$p_{iG} = \frac{1}{\left(1 + \sum_{h=1}^{G-1} e^{x_i^T \beta_h}\right)} \qquad \text{(A3)}$$

wherein

P_{ig} is the probability that the training sample with input data X_i will be in sample class g;
$x_i^T \beta_g$ is a linear combination generated from input data from training sample i with component weights β_g;
x_i^T is the components for the i^{th} Row of X and β_g is a set of component weights for sample class g;

Typically, as discussed above, the component weights are estimated in a manner which takes into account the apriori assumption that most of the component weights are zero.

In one embodiment, components weights βg in equation (A2) are estimated in a manner whereby most of the values are zero, yet the samples can still be accurately classified.

In one embodiment, the component weights are estimated by maximising the posterior distribution of the weights given the data in the Bayesian model referred to above.

Preferably, the component weights are estimated by

(a) specifying a hierarchical prior for the component weights $\beta_1,...,\beta_{G-1}$; and

(b) specifying a likelihood function for the input data;

(c) determining the posterior distribution of the weights given the data using (A5); and

(d) determining component weights which maximise the posterior distribution.

In one embodiment, the hierarchical prior specified for the parameters $\beta_1,...,\beta_{G-1}$ is of the form:

$$P(\beta_1...,\beta_{G-1}) = \int_{\tau^2} \prod_{g=1}^{G-1} P(\beta_g|\tau_g^2) P(\tau_g^2) d\tau^2 \qquad \text{(A4)}$$

where $\beta^T = (\beta_1^T,...\beta_{G-1}^T)$, $\tau^T = (\tau_1^T,...,\tau_{G-1}^T)$, $p(\beta_g|\tau_g')$ is $N(0,\text{diag}\{\tau_g'\})$ and $p(\tau_g^2)$ is a suitable prior.

In one embodiment, $p(\tau_g^2) = \prod_{i=1}^{n} p(\tau_{ig}^2)$ where $p(\tau_{ig}^2)$ is a prior

wherein $t_{ig}^2 = 1/\tau_{ig}^2$ has an independent gamma distribution.

In another embodiment, $p(\tau_{ig}^2)$ is a prior wherein τ_{ig}^2 has an independent gamma distribution.

In one embodiment, the likelihood function is $L(\underset{\sim}{y}|\beta_1,...,\beta_{G-1})$ of the form in equation (8) and the posterior distribution of β and $\underset{\sim}{\tau}$ given y is

$$p(\beta\tau^2|y) \alpha L(y|\beta) p(\beta|\tau^2) p(\tau^2) \qquad \text{(A5)}$$

In one embodiment, the likelihood function has a first and second derivative.

In one embodiment, the first derivative is determined from the following algorithm:

$$\frac{\partial \log L}{\partial \beta g} = X^T\left(\underset{\sim}{e}_g - p_g\right), \qquad g = 1,...,G-1 \qquad \text{(A6)}$$

wherein $\underset{\sim}{e}_g^T = \left(e_{ig}, i = 1,...,n\right)$, $p_g^T = \left(p_{ig}, i = 1,...,n\right)$ are vectors indicating membership of sample class g and probability of class g respectively.

In one embodiment, the second derivative is determined from the following algorithm:

$$\frac{\partial^2 \log L}{\partial \beta_g \partial \beta_h} = -X^T diag\left\{\delta_{hg} p_g - p_h p_g\right\} X \qquad \text{(A7)}$$

where δ_{hg} is 1 if h equals g and zero otherwise.

Equation A6 and equation A7 may be derived as follows:

(a) Using equations (A1), (A2) and (A3), the likelihood function of the data can be written as:

$$L = \prod_{i=1}^{n}\left(\prod_{g=1}^{G-1}\left\{\frac{e^{x_i^T\beta_g}}{\left(1+\sum_{g=1}^{G-1}e^{x_i^T\beta_g}\right)}\right\}^{e_{ik}}\left\{\frac{1}{1+\sum_{h=1}^{G-1}e^{x_i^T\beta_h}}\right\}^{e_{iG}}\right) \qquad \text{(A8)}$$

(b) Taking logs of equation (A6) and using the fact that $\sum_{h=1}^{G} e_{ih} = 1$ for all i gives:

$$l\log L = \sum_{i=1}^{n}\left(\sum_{g=1}^{G-1} e_{ig} x_i^T \beta_g - \log\left(1+\sum_{g=1}^{G-1} e^{x_i^T\beta_g}\right)\right) \qquad \text{(A9)}$$

(c) Differentiating equation (A8) with respect to βg gives

$$\frac{\partial \log L}{\partial \beta g} = X^T\left(\underset{\sim}{e}_g - p_g\right), \qquad g = 1,\ldots,G-1 \tag{A10}$$

whereby $\underset{\sim}{e}_g^T = \left(e_{ig}, i = 1,n\right)$, $p_g^T = \left(p_{ig}, i = 1,n\right)$ are vectors indicating membership of sample class g and probability of class g respectively.

(d) The second derivative of equation (9) has elements

$$\frac{\partial^2 \log L}{\partial \beta_g \partial \beta_h} = -X^T diag\left\{\delta_{hg} p_g - p_h p_g\right\} X \tag{A11}$$

where

$$\delta_{hg} = \begin{cases} 1, & h = g \\ 0, & \text{otherwise} \end{cases}$$

Component weights which maximise the posterior distribution of the likelihood function may be specified using an EM algorithm comprising an E step and an M step.

In conducting the EM algorithm, the E step preferably comprises the step computing a term of the form:

$$\begin{aligned} \mathrm{P} &= \sum_{g=1}^{G-1} \sum_{i=1}^{n} \mathrm{E}\{\beta_{ig}^2/\tau_{ig}^2 \mid \hat{\beta}_{ig}\} \\ &= \sum_{g=1}^{G-1} \sum_{i=1}^{n} \beta_{ig}^2/\hat{d}_{ig}^2 \end{aligned} \tag{A11a}$$

where $\hat{d}_{ig} = E\{1/\tau_{ig}^2 \mid \hat{\beta}_{ig}\}^{-0.5}$, $\hat{d}_g = (\hat{d}_{1g}, \hat{d}_{2g}, \ldots, \hat{d}_{pg})^T$ and $\hat{d}_{ig} = 1/\hat{d}_{ig} = 0$ if $\hat{\beta}_{ig} = 0$.

Preferably, equation (11a) is computed by calculating the conditional expected value of $t_{ig}^2 = 1/\tau_{ig}^2$ when $p\left(\beta_{ig} \middle| \tau_{ig}^2\right)$ is $N\left(0, \tau_{ig}^2\right)$ and $p\left(\tau_{ig}^2\right)$ has a specified prior distribution.

Explicit formulae for the conditional expectation will be presented later.

Typically, the EM algorithm comprises the steps:

(a) performing an E step by calculating the conditional expected value of the posterior distribution of component weights using the function:

$$Q = Q(\gamma \mid y, \hat{\gamma}) = \log L - \frac{1}{2}\sum_{g=1}^{G-1} \gamma_g^T \, diag\left\{d_g(\hat{\gamma}_g)\right\}^{-2} \gamma_g \qquad \text{(A12)}$$

where $x_i^T \beta_g = x_i^T P_g \gamma_g$ in equation (8), $d(\hat{\gamma}_g) = P_g^T \hat{d}_g$, and $\hat{d}_g$ is defined as in equation (11a) evaluated at $\hat{\beta}_g = P_g \hat{\gamma}_g$. Here P_g is a matrix of zeroes and ones derived from the identity matrix such that $P_g{}^T \beta_g$ selects non-zero elements of β_g which are denoted by γ_g.

(b) performing an M step by applying an iterative procedure to maximise Q as a function of γ whereby:

$$\gamma^{t+1} = \gamma^t - \alpha^t \left(\frac{\partial^2 Q}{\partial \gamma^2}\right)^{-1} \left(\frac{\partial Q}{\partial \gamma}\right) \qquad \text{(A13)}$$

where α^t is a step length such that $0 \le \alpha^t \le 1$; and $\gamma = (\gamma_g, g=1,\ldots..,G-1)$.

Equation (A12) may be derived as follows:

Calculate the conditional expected value of (A5) given the observed data y and a set of parameter estimates $\hat{\beta}$.

$$Q = Q\left(\beta \middle| \underset{\sim}{y}, \hat{\beta}\right) = \mathrm{E}\left\{\log p\left(\underset{\sim}{\beta}, \underset{\sim}{\tau} | y\right) \middle| y, \hat{\beta}\right\}$$

Consider the case when components of β (and $\hat{\beta}$) are set to zero i.e for $g = 1,\ldots, G-1$, $\beta_g = P_g \gamma_g$ and $\ddot{\beta}_g = P_g \hat{\gamma}_g$.

Ignoring terms not involving γ and using (A4), (A5), (A9) we get:

$$Q = \log L - \frac{1}{2}\sum_{g=1}^{G-1}\sum_{i=1}^{n} \mathrm{E}\left\{\frac{\gamma_{ig}^2}{\tau_{ig}^2} \middle| y, \hat{\underset{\sim}{\gamma}}\right\}$$

$$Q = \log L - \frac{1}{2}\sum_{g=1}^{G-1}\sum_{i=1}^{n} \gamma_{ig}^2 \mathrm{E}\left\{\frac{1}{\tau_{ig}^2} \middle| y, \hat{\underset{\sim}{\gamma}}\right\}$$

$$= \log L - \frac{1}{2}\sum_{g=1}^{G-1} \gamma_g^T \, diag\left\{d_g(\hat{\gamma}_g)\right\}^{-2} \gamma_g \qquad \text{(A14)}$$

where $x_i^T \beta_g = x_i^T P_g \hat{\gamma}_g$ in (A8), $d_g(\hat{\gamma}_g) = P_g^T \hat{d}_g$ where $\hat{d}$ is defined as in equation (A11a) evaluated at $\ddot{\beta}_g = P_g \hat{\gamma}_g$.

Note that the conditional expectation can be evaluated from first principles given (A4). Some explicit expressions are given later.

The iterative procedure may be derived as follows:
To obtain the derivatives required in (11), first note that from (A8), (A9) and (A10) writing $d(\hat{\gamma}) = \{d_g(\hat{\gamma}_g), g = 1,\ldots, G-1\}$, we get

$$\frac{\partial Q}{\partial \gamma} = \left(\frac{\partial \beta}{\partial \gamma}\right)\frac{\partial \log L}{\partial \beta} - diag\left\{d(\hat{\gamma})\right\}^{-2}\gamma$$

$$= \begin{bmatrix} X_1^T(e_1 - p_1) \\ \vdots \\ X_{G-1}^T(e_{G-1} - p_{G-1}) \end{bmatrix} - diag\left\{d(\hat{\underset{\sim}{\gamma}})\right\}^{-2}\gamma \qquad \text{(A15)}$$

and

$$\frac{\partial^2 Q}{\partial \underset{\sim}{\gamma}^2} = \left(\frac{\partial \beta}{\partial \gamma}\right) \frac{\partial^2 \log L}{\partial_\beta^2} \left(\frac{\partial \beta}{\partial \gamma}\right)^T - diag\left\{d(\hat{\gamma})\right\}^{-2}$$

$$= -\left\{ \begin{pmatrix} X_1^T \Delta_{11} X_1 & \cdots & X_1^T \Delta_{1G-1} X_{G-1} \\ \vdots & & \vdots \\ X_{G-1} \Delta_{G-11} X_1 & & X_{G-1} \Delta_{G-1G-11} X_{G-1} \end{pmatrix} + diag\left\{d(\hat{\gamma})\right\}^{-2} \right\} \quad \text{(A16)}$$

where

$$\Delta_{gh} = diag\left\{\delta_{gh} p_g - p_g p_h\right\},$$

$$\delta_{gh} = \begin{cases} 1, & g = h \\ 0, & \text{otherwise} \end{cases}$$

$$d(\hat{\gamma}) = (d(\hat{\gamma}_g), g = 1, \ldots, G-1)$$

and

$$X_g^T = P_g^T X^T, g = 1, \ldots G-1 \quad \text{(A17)}$$

In a preferred embodiment, the iterative procedure may be simplified by using only the block diagonals of equation (A16) in equation (A13). For $g = 1, \ldots G-1$, this gives:

$$\gamma_g^{t+1} = \gamma_g^t + \alpha^t \left\{ X_g^T \Delta_{gg} X_g + diag\left\{d_g(\hat{\gamma}_g)\right\}^{-2} \right\}^{-1} \left\{ X_g^T \left(e_g - p_g\right) - diag\left\{d_g(\hat{\gamma}_g)\right\} \gamma_g^t \right\} -$$

(A18)

Rearranging equation (A18) leads to

$$\gamma_g^{t+1} = \gamma_g^t + \alpha^t diag\left\{d_g(\hat{\gamma}_g)\right\} \left(Y_g^T \Delta_{gg} Y_g + I\right)^{-1} \left\{ Y_g^T \left(e_g - p_g\right) - diag\left\{d_g(\hat{\gamma}_g)\right\}^{-1} \gamma_g^t \right\}^{-1}$$

(A19)

where

$$Y_g^T = diag\left\{d_g(\hat{\gamma}_g)\right\} X_g^T$$

Writing $p(g)$ for the number of columns of Y_g, (A19) requires the inversion of a $p(g) \times p(g)$ matrix which may be quite large. This can be reduced to an $n \times n$ matrix for $p(g) > n$ by noting that:

$$\begin{aligned}\left(Y_g^T \Delta_{gg} Y_g + I\right)^{-1} &= I - Y_g^t \left(Y_g Y_g^T + \Delta_{gg}^{-1}\right)^{-1} Y_g \\ &= I - Z_g^T \left(Z_g Z_g^T + I_n\right)^{-1} Z_g\end{aligned} \qquad \text{(A20)}$$

where $Z_g = \Delta_{gg}^2 Y_g$. Preferably, (A19) is used when $p(g) > n$ and (A19) with (A20) substituted into equation (A19) is used when $p(g) \leq n$.

Note that when τ_{ig}^2 has a Jeffreys prior we have:

$$E\{t_{ig}^2 \mid \hat{\beta}_{ig}\} = 1/\hat{\beta}_{ig}^2$$

In one embodiment, $t_{ig}^2 = 1/\tau_{ig}^2$ has an independent *gamma distribution* with scale parameter b>0 and shape parameter k>0 so that the density of t_{ig}^2 is:

$$\gamma(t_{ig}^2, b, k) = b^{-1} (t_{ig}^2/b)^{k-1} \exp(-t_{ig}^2/b)/\Gamma(k)$$

Omitting subscripts to simplify the notation, it can be shown that

$$E\{ t^2 \mid \beta \} = (2k+1)/(2/b + \beta^2) \qquad \text{(A21)}$$

as follows:

Define

$$I(p,b,k) = \int_0^\infty (t^2)^p \, t \exp(-0.5\beta^2 t^2) \gamma(t^2, b, k) dt^2$$

then

$$I(p,b,k)=b^{p+0.5}\{\Gamma(p+k+0.5)/\Gamma(k)\}(1+0.5b\beta^2)^{-(p+k+0.5)}$$

Proof

Let $s=\beta^2/2$ **then**

$$I(p,b,k)=b^{p+0.5}\int_0^\infty (t^2/b)^{p+0.5}\exp(-st^2)\gamma(t^2,b,k)dt^2$$

Now using the substitution u = t^2/b we get

$$I(p,b,k)=b^{p+0.5}\int_0^\infty (u)^{p+0.5}\exp(-sub)\gamma(u,1,k)du$$

Now let s′=bs and substitute the expression for $\gamma(u,1,k)$**. This gives**

$$I(p,b,k)=b^{p+0.5}\int_0^\infty \exp(-(s'+1)u)u^{p+k+0.5-1}du\,/\,\Gamma(k)$$

Looking up a table of Laplace transforms, eg Abramowitz and Stegun, then gives the result.

The conditional expectation follows from

$$E\{t^2\mid\beta\}=I(1,b,k)/I(0,b,k)$$
$$=(2k+1)/(2/b+\beta^2)$$

As k tends to zero and b tends to infinity we get the equivalent result using Jeffreys prior. For example, for k=0.005 and b=2*10^5

$$E\{t^2\mid\beta\}=(1.01)/(10^{-5}+\beta^2)$$

Hence we can get arbitrarily close to the Jeffreys prior with this proper prior.

The algorithm for this model has

$$\hat{d} = E\{t^2 | \hat{\beta}\}^{-0.5}$$
$$= E\{\frac{1}{\tau^2} | \hat{\beta}\}^{-0.5}$$

where the expectation is calculated as above.

In another embodiment, τ_{ig}^2 has an independent *gamma distribution* with scale parameter b>0 and shape parameter k>0. It can be shown that

$$E\{\tau_{ig}^{-2} | \beta_{ig}\} = \frac{\int_0^\infty u^{k-3/2-1} \exp(-(\gamma_{ig}/u + u))\, du}{b\int_0^\infty u^{k-1/2-1} \exp(-(\gamma_{ig}/u + u))\, du}$$
$$= \sqrt{\frac{2}{b}} \frac{1}{|\beta_{ig}|} \frac{K_{3/2-k}(2\sqrt{\gamma_{ig}})}{K_{1/2-k}(2\sqrt{\gamma_{ig}})} \qquad \text{(A22)}$$
$$= \frac{1}{|\beta_{ig}|^2} \frac{(2\sqrt{\gamma_{ig}}) K_{3/2-k}(2\sqrt{\gamma_{ig}})}{K_{1/2-k}(2\sqrt{\gamma_{ig}})}$$

where $\gamma_{ig} = \beta_{ig}^2/2b$ and K denotes a modified Bessel function. For k=1 in equation (A22)

$$E\{\tau_{ig}^{-2} | \beta\} = \sqrt{2}/b(1/|\beta_{ig}|)$$

For K=0.5 in equation (A22)

$$E\{\tau_{ig}^{-2} | \beta_{ig}\} = \sqrt{2}/b(1/|\beta_{ig}|)\{K_1(2\sqrt{\gamma_{ig}})/K_0(2\sqrt{\gamma_{ig}})\}$$

or equivalently

$$E\{\tau_{ig}^{-2} | \beta_i\} = (1/|\beta_{ig}|^2)\{2\sqrt{\gamma_i} K_1(2\sqrt{\gamma_{ig}})/K_0(2\sqrt{\gamma_{ig}})\}$$

Proof of (A.1)

From the definition of the conditional expectation, writing $\gamma = \beta^2/2b$, we get

$$E\{\tau^{-2}|\beta\}=\frac{\int_0^\infty \tau^{-2}\tau^{-1}\exp(-\gamma\tau^{-2})b^{-1}(\tau^{-2}/b)^{k-1}\exp(\tau^{-2}/b)d\tau^2}{\int_0^\infty \tau^{-1}\exp(-\gamma\tau^{-2})b^{-1}(\tau^{-2}/b)^{k-1}\exp(\tau^{-2}/b)d\tau^2}$$

Rearranging, simplifying and making the substitution $u=\tau^2/b$ gives the first equation in (A22).

The integrals in (22) can be evaluated by using the result

$$\int_0^\infty x^{-b-1}\exp\left[-\left(x+\frac{a^2}{x}\right)\right]dx=\frac{2}{a^b}K_b(2a)$$

where K denotes a modified Bessel function, see Watson(1966).

Examples of members of this class are k=1 in which case

$$E\{\tau_{ig}^{-2}|\beta_{ig}\}=\sqrt{2/b}(1/|\beta_{ig}|)$$

which corresponds to the prior used in the Lasso technique, Tibshirani(1996). See also Figueiredo(2001).

The case k=0.5 gives

$$E\{\tau_{ig}^{-2}|\beta_{ig}\}=\sqrt{2/b}(1/|\beta_{ig}|)\{K_1(2\sqrt{\gamma_{ig}})/K_0(2\sqrt{\gamma_{ig}})\}$$

or equivalently

$$E\{\tau_{ig}^{-2}|\beta_{ig}\}=(1/|\beta_{ig}|^2)\{2\sqrt{\gamma_i}K_1(2\sqrt{\gamma_{ig}})/K_0(2\sqrt{\gamma_{ig}})\}$$

where K_0 and K_1 are modified Bessel functions, see Abramowitz and Stegun(1970). Polynomial approximations for evaluating these Bessel functions can be found in Abramowitz and Stegun(1970, p379). The expressions above demonstrate the connection with the Lasso model and the Jeffreys prior model.

It will be appreciated by those skilled in the art that as k tends to zero and b tends to infinity the prior tends to a Jeffreys improper prior.

In one embodiment, the priors with $0< k \leq 1$ and $b>0$ form a class of priors which might be interpreted as penalising non zero coefficients in a manner which is between the Lasso prior and the specification using Jeffreys hyper prior.

The hyperparameters b and k can be varied to control the number of components selected by the method. As k tends to zero for fixed b the number of components selected can be decreased and conversely as k tends to 1 the number of selected components can be increased.

In a preferred embodiment, the EM algorithm is performed as follows:

1. Set n=0 , $P_g = I$ and choose an initial value for $\hat{\gamma}^0$.Choose a value for b and k in equation (A22). For example b=1e7 and k=0 gives the Jeffreys prior model to a good degree of approximation. This is done by ridge regression of $\log(p_{ig}/p_{iG})$ on x_i where p_{ig} is chosen to be near one for observations in group g and a small quantity >0 otherwise - subject to the constraint of all probabilities summing to one.

2. Do the E step i.e evaluate $Q = Q\left(\gamma \middle| \underset{\sim}{y}, \hat{\gamma}^n\right)$. Note that this also depends on the values of k and b.

3. Set t=0. For $g = 1,...,G-1$ calculate:
   a) $\delta_g^t = \gamma_g^{t+1} - \gamma_g^t$ using (A19) with (A20) substituted into (A19) when $p(g) \geq n$.
   (b) Writing $\delta^t = \left(\delta_g^t, g = 1,...,G-1\right)$ Do a line search to find the value of α^t in $\underset{\sim}{\gamma}^{t+1} = \underset{\sim}{\gamma}^t + \alpha^t \underset{\sim}{\delta}^t$ which maximises (or simply increases) (12) as a function of α^t.

c) set $\underset{\sim}{\gamma}^{t+1} = \underset{\sim}{\gamma}^{t}$ and t=t+1

Repeat steps (a) and (b) until convergence.

This produces γ^{*n+1} say which maximises the current Q function as a function of γ .

For $g = 1,...G-1$ determine $S_g = \left\{ j : \left| \gamma_{jg}^{*n+1} \right| \leq \varepsilon \max_k \left| \gamma_{kg}^{*n+1} \right| \right\}$

Where $\varepsilon \ll 1$, say 10^{-5}. Define P_g so that $\beta_{ig} = 0$ for $i \in S_g$ and

$$\hat{\gamma}_g^{n+1} = \left\{ \gamma_{jg}^{*n+1}, j \notin S_g \right\}$$

This step eliminates variables with small coefficients from the model.

4. Set n=n+1 and go to 2 until convergence.

A second embodiment relating to an ordered cat logistic regression will now be described.

B. Ordered categories model

The method of this embodiment may utilise the training samples in order to identify a subset of components which can be used to determine whether a test sample belongs to a particular class. For example, to identify genes for assessing a tissue biopsy sample using microarray analysis, microarray data from a series of samples from tissue that has been previously ordered into classes of increasing or decreasing disease severity such as normal tissue, benign tissue, localised tumour and metastasised tumour tissue are used as training samples to identify a subset of components which is capable of indicating the severity of disease associated with the training samples. The subset of components can then be subsequently used to determine whether previously unclassified test samples can be

classified as normal, benign, localised tumour or metastasised tumour. Thus, the subset of components is diagnostic of whether a test sample belongs to a particular class within an ordered set of classes. It will be apparent that once the subset of components have been identified, only the subset of components need be tested in future diagnostic procedures to determine to what ordered class a sample belongs.

The method of the invention is particularly suited for the analysis of very large amounts of data. Typically, large data sets obtained from test samples is highly variable and often differs significantly from that obtained from the training samples. The method of the present invention is able to identify subsets of components from a very large amount of data generated from training samples, and the subset of components identified by the method can then be used to classifying test samples even when the data generated from the test sample is highly variable compared to the data generated from training samples belonging to the same class. Thus, the method of the invention is able to identify a subset of components that are more likely to classify a sample correctly even when the data is of poor quality and/or there is high variability between samples of the same ordered class.

The components are "predictive" for that particular ordered class. Essentially, from all the data which is generated from the system, the method of the present invention enables identification of a relatively small number of components which can be used to classify the training data. Once those components have been identified by this method, the components can be used in future to classify test samples. The method of the present invention preferably utilises a statistical method to eliminate components that are not required to correctly classify the sample into a class that is a member of an ordered class.

In the following there are N samples, and vectors such as y, z and μ have components y_i, z_i and μ_i for i = 1, ..., N. Vector multiplication and division is defined componentwise and diag{ · } denotes a diagonal matrix whose diagonals are equal to the argument. We also use || · || to denote Euclidean norm.

Preferably, there are N observations y_i^* where y_i^* takes integer values 1,...,G. The values denote classes which are ordered in some way such as for example severity of disease. Associated with each observation there is a set of covariates (variables, e.g gene expression values) arranged into a matrix X^* with n row and p columns wherein n is the samples and p the components. The notation x_i^{*T} denotes the i^{th} row of X^*. Individual (sample) i has probabilities of belonging to class k given by $\pi_{ik} = \pi_k(x_i^*)$.
Define cumulative probabilities

$$\gamma_{ik} = \sum_{g=1}^{k} \pi_{ik} \quad , \qquad k = 1,...,G$$

Note that γ_{ik} is just the probability that observation i belongs to a class with index less than or equal to k. Let C be a N by p matrix with elements c_{ij} given by

$$c_{ij} = \begin{cases} 1, \text{ if observation i in class j} \\ 0, \text{ otherwise} \end{cases}$$

and let R be an n by P matrix with elements r_{ij} given by

$$r_{ij} = \sum_{g=1}^{j} c_{ig}$$

These are the cumulative sums of the columns of C within rows.

For independent observations (samples) the likelihood of the data may be written as:

$$l=\prod_{i=1}^{N}\prod_{k=1}^{G-1}\left(\frac{\gamma_{ik}}{\gamma_{ik+1}}\right)^{r_{ik}}\left(\frac{\gamma_{ik+1}-\gamma_{ik}}{\gamma_{ik+1}}\right)^{r_{ik+1}-r_{ik}}$$

and the log likelihood L may be written as:

$$L=\sum_{i=1}^{N}\sum_{j=1}^{G-1}r_{ik}\log\left(\frac{\gamma_{ik}}{\gamma_{ik+1}}\right)+(r_{ik+1}-r_{ik})\log\left(\frac{\gamma_{ik+1}-\gamma_{ik}}{\gamma_{ik+1}}\right)$$

The continuation ratio model may be adopted here as follows:

$$\operatorname{logit}\left(\frac{\gamma_{ik+1}-\gamma_{ik}}{\gamma_{ik+1}}\right)=\operatorname{logit}\left(\frac{\pi_{ik}}{\gamma_{ik+1}}\right)=\theta_k+x_i^T\beta^*$$

for k = 2,...,G , see McCullagh and Nelder(1989) and McCullagh(1980) and the discussion therein. Note that

$$\operatorname{logit}\left(\frac{\gamma_{ik+1}-\gamma_{ik}}{\gamma_{ik+1}}\right)=-\operatorname{logit}\left(\frac{\gamma_{ik}}{\gamma_{ik+1}}\right).$$

The likelihood is equivalent to a logistic regression likelihood with response vector y and covariate matrix X

$$y=vec\{R\}$$
$$X=\left[B_1^T B_2^T \ldots B_N^T\right]^T$$
$$B_i=\left[I_{G-1}\,|\,1_{G-1}x_i^{*T}\right]$$

where I_{G-1} is the G-1 by G-1 identity matrix and 1_{G-1} is a G-1 by 1 vector of ones.
Here vec{ } takes the matrix and forms a vector row by row.

Typically, as discussed above, the component weights are estimated in a manner which takes into account the a priori assumption that most of the component weights are zero.

Following Figueiredo(2001), in order to eliminate redundant variables (covariates), a prior is specified for the

parameters β^* by introducing a p x 1 vector of hyperparameters.

Preferably, the prior specified for the component weights is of the form

$$p(\beta^*) = \int_{\tau^2} p(\beta^* | \nu^2) p(\nu^2) d\nu^2 \qquad \text{(B1)}$$

where $p(\beta^* | \nu^2)$ is $N(0, \text{diag}\{\nu^2\})$ and $p(\nu^2)$ is a suitably chosen hyperprior. For example, $p(\nu^2) \alpha \prod_{i=1}^{n} p(\nu_i^2)$ is a suitable form of Jeffreys prior.

In another embodiment, $p(\nu_i^2)$ is a prior wherein $t_i^2 = 1/\nu_i^2$ has an independent gamma distribution.

In another embodiment, $p(\nu_i^2)$ is a prior wherein ν_i^2 has an independent gamma distribution.

The elements of theta have a non informative prior.

Writing $L(\underset{\sim}{y} | \beta^* \theta)$ for the likelihood function, in a Bayesian framework the posterior distribution of β, θ and ν given y is

$$p(\beta^* \varphi \nu | y) \alpha L(y | \beta^* \varphi) p(\beta^* | \nu) p(\nu) \qquad \text{(2)}$$

By treating ν as a vector of missing data, an iterative algorithm such as an EM algorithm (Dempster et al, 1977) can be used to maximise (2) to produce maximum a posteriori estimates of β and θ. The prior above is such that the

maximum a posteriori estimates will tend to be sparse i.e. if a large number of parameters are redundant, many components of β^* will be zero.

Preferably $\beta^T=(\theta^T,\beta^{*T})$ in the following:

For the ordered categories model above it can be shown that

$$\frac{\partial L}{\partial \beta}=X^t(y-\mu) \qquad (11)$$

$$E\{\frac{\partial^2 L}{\partial \beta^2}\}=-X^t\mathrm{diag}\{\mu(1-\mu)\}X \qquad (12)$$

Where $\mu_i=\exp(x_i^T\beta)/(1+\exp(x_i^T\beta))$ and $\beta^T=(\theta_2,...,\theta_G,\beta^{*T})$.

The iterative procedure for maximising the posterior distribution of the components and component weights is an EM algorithm, such as, for example, that described in Dempster et al, 1977. Preferably, the EM algorithm is performed as follows:

1. Chose a hyperprior and values b and k for its parameters. Set n=0, $S_0 = \{1,2,...,p\}$, $\phi^{(0)}$, and $\varepsilon = 10^{-5}$ (say). Set the regularisation parameter κ at a value much greater than 1, say 100. This corresponds to adding $1/\kappa^2$ to the first G-1 diagonal elements of the second derivative matrix in the M step below.

If $p \leq N$ compute initial values β^* by

$$\beta^*=(X^tX+\lambda I)^{-1}X^Tg(y+\zeta) \qquad (B2)$$

and if $p > N$ compute initial values β^* by

$$\beta^*=\frac{1}{\lambda}(I-X^T(XX^T+\lambda I)^{-1}X)X^Tg(y+\zeta) \qquad (B3)$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$ and ζ is small and chosen so that the link function $g(z)=\log(z/(1-z))$ is well defined at $z = y+\zeta$.

2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

and let P_n be a matrix of zeroes and ones such that the nonzero elements $\gamma^{(n)}$ of $\beta^{(n)}$ satisfy

$$\gamma^{(n)} = P_n^T \beta^{(n)} \quad , \quad \beta^{(n)} = P_n \gamma^{(n)}$$
$$\gamma = P_n^T \beta \quad , \quad \beta = P_n \gamma$$

Define $w_\beta = (w_{\beta i}, i = 1, p)$, such that

$$w_{\beta i} = \begin{cases} 1, & i \geq G \\ 0, & \text{otherwise} \end{cases}$$

and let $w_\gamma = P_n w_\beta$

3. Perform the E step by calculating

$$\begin{aligned} Q(\beta \mid \beta^{(n)}, \varphi^{(n)}) &= E\{ \log p(\beta, \varphi, \nu \mid y) \mid y, \beta^{(n)}, \phi^{(n)} \} \\ &= L(y \mid \beta, \phi^{(n)}) - 0.5 \, (\| (\beta^* w_\beta) / \hat{d}^{(n)} \|^2) \end{aligned} \qquad (15)$$

where L is the log likelihood function of y and $\hat{d}_{ig} = E\{1/\tau_{ig}^2 \mid \hat{\beta}_{ig}\}^{-0.5}$, $\hat{d}_g = (\hat{d}_{1g}, \hat{d}_{2g}, ..., \hat{d}_{pg})^T$ and for convenience we define $\hat{d}_{ig} = 1/\hat{d}_{ig} = 0$ if $\hat{\beta}_{ig} = 0$. . Using $\beta = P_n \gamma$ and $\beta^{(n)} = P_n \gamma^{(n)}$ (15) can be written as

$$Q(\gamma \mid \gamma^{(n)}, \phi^{(n)}) = L(y \mid P_n \gamma, \phi^{(n)}) - 0.5 \, (\| (\gamma^* w_\gamma) / d(\gamma^{(n)}) \|^2) \qquad (B4)$$

with $d(\gamma^{(n)}) = P_n^T d^{(n)}$ evaluated at $\beta^{(n)} = P_n \gamma^{(n)}$.

4. Do the M step. This can be done with Newton Raphson iterations as follows. Set $\gamma_0 = \gamma^{(n)}$ and for r=0,1,2,... γ_{r+1}

$= \gamma_r + \alpha_r \delta_r$ where α_r is chosen by a line search algorithm to ensure $Q(\gamma_{r+1} | \gamma^{(n)}, \phi^{(n)}) > Q(\gamma_r | \gamma^{(n)}, \phi^{(n)})$.

For $p \leq N$ use

$$\delta_r = \Delta(d^*(\gamma^{(n)}))[Y_n^T V_r^{-1} Y_n + I]^{-1}(Y_n^T z_r - \frac{w_\gamma \gamma_r}{d^*(\gamma^{(n)})}) \quad \text{(B5)}$$

where

$$d^*(\gamma^{(n)}) = \begin{cases} d(\gamma^{(n)}), & i \geq G \\ \kappa, & \text{otherwise} \end{cases}$$

$$Y_n^T = \Delta(d^*(\gamma^{(n)}))P_n^T X^T$$

$$V_r^{-1} = \text{diag}\{\mu_r(1-\mu_r)\}$$

$$z_r = (y-\mu_r)$$

and $\mu_r = \exp(XP_n\gamma_r)/(1+\exp(XP_n\gamma_r))$.

For $p > N$ use

$$\delta_r = \Delta(d^*(\gamma^{(n)}))[I - Y_n^T(Y_n Y_n^T + V_r)^{-1} Y_n](Y_n^T z_r - \frac{w_\gamma \gamma_r}{d^*(\gamma^{(n)})}) \quad \text{(B6)}$$

with V_r and z_r defined as before.

Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g
$|| \gamma_r - \gamma_{r+1} || < \varepsilon$ (for example 10^{-5}).

5. Define $\beta^* = P_n\gamma^*$, $S_{n+1} = \{i \geq G: |\beta_i| > \max_{j \geq G}(|\beta_j| * \epsilon_1)\}$ where ε_1 is a small constant, say 1e-5. Set n=n+1.

6. Check convergence. If $|| \gamma^* - \gamma^{(n)} || < \varepsilon_2$ where ε_2 is suitably small then stop, else go to step 2 above.

Recovering the probabilities
Once we have obtained estimates of the parameters β are obtained, calculate

$$a_{ik}=\frac{\hat{\pi}_{ik}}{\hat{\gamma}_{ik}}$$

for i=1,…,N and k = 2,…,G.

Preferably, to obtain the probabilities we use the recursion

$$\pi_{iG}=a_{iG}$$

$$\pi_{ik-1}=\left(\frac{a_{ik-1}}{a_{ik}}\right)(1-a_{ik})\pi_{ik}$$

and the fact that the probabilities sum to one, for i = 1,…,N.

In one embodiment the covariate matrix X with rows x_i^T can be replaced by a matrix K with ij^{th} element k_{ij} and $k_{ij} = \kappa(x_i - x_j)$ for some kernel function κ. This matrix can also be augmented with a vector of ones. Some example kernels are given in Table 1 below, see Evgeniou et al(1999).

Kernel function	Formula for $\kappa(x - y)$
Gaussian radial basis function	$\exp(- \|\| x - y \|\|^2 / a)$, a>0
Inverse multiquadric	$(\|\| x - y \|\|^2 + c^2)^{-1/2}$
multiquadric	$(\|\| x - y \|\|^2 + c^2)^{½}$
Thin plate splines	$\|\| x - y \|\|^{2n+1}$ $\|\| x - y \|\|^{2n} \ln(\|\| x - y \|\|)$
Multi layer perceptron	$\tanh(x \cdot y-\theta)$, for suitable θ
Ploynomial of degree d	$(1 + x \cdot y)^d$
B splines	$B_{2n+1}(x - y)$
Trigonometric polynomials	$\sin((d+1/2)(x-y))/\sin((x-y)/2)$

Table 1: Examples of kernel functions

In Table 1 the last two kernels are preferably one dimensional i.e. for the case when X has only one column. Multivariate versions can be derived from products of these kernel functions. The definition of B_{2n+1} can be found in De Boor(1978). Use of a kernel function results in mean values

which are smooth (as opposed to transforms of linear) functions of the covariates X. Such models may give a substantially better fit to the data.

A third embodiment relating to generalised linear models will now be described.

C. Generalised Linear Models

The method of this embodiment utilises the training samples in order to identify a subset of components which can predict the characteristic of a sample. Subsequently, knowledge of the subset of components can be used for tests, for example clinical tests to predict unknown values of the characteristic of interest. For example, a subset of components of a DNA microarray may be used to predict a clinically relevant characteristic such as, for example, a blood glucose level, a white blood cell count, the size of a tumour, tumour growth rate or survival time.

In this way, the present invention identifies preferably a relatively small number of components which can be used to predict a characteristic for a particular sample. The selected components are "predictive" for that characteristic. By appropriate choice of hyperparameters in the hyper prior the algorithm can be made to select subsets of varying sizes. Essentially, from all the data which is generated from the system, the method of the present invention enables identification of a small number of components which can be used to predict a particular characteristic. Once those components have been identified by this method, the components can be used in future to predict the characteristic for new samples. The method of the present invention preferably utilises a statistical method to eliminate components that are not required to correctly predict the characteristic for the sample.

The inventors have found that component weights of a linear combination of components of data generated from the training samples can be estimated in such a way as to eliminate the components that are not required to predict a characteristic for a training sample. The result is that a subset of components are identified which can correctly predict the characteristic for samples in the training set. The method of the present invention thus permits identification from a large amount of data a relatively small number of components which are capable of correctly predicting a characteristic for a training sample, for example, a quantity of interest.

The characteristic may be any characteristic of interest. In one embodiment, the characteristic is a quantity or measure. In another embodiment, they may be the index number of a group, where the samples are grouped into two sample groups (or "classes") based on a pre-determined classification. The classification may be any desired classification by which the training samples are to be grouped. For example, the classification may be whether the training samples are from a leukemia cell or a healthy cell, or that the training samples are obtained from the blood of patients having or not having a certain condition, or that the training samples are from a cell from one of several types of cancer as compared to a normal cell. In another embodiment the characteristic may be a censored survival time, indicating that particular patients have survived for at least a given number of days. In other embodiments the quantity may be any continuously variable characteristic of the sample which is capable of measurement, for example blood pressure.

In one embodiment, the data may be a quantity y_i, where $i \in \{1,...,N\}$. We write the nx1 vector with elements y_i as y. We define a p x 1 parameter vector β of component weights (many of which are expected to be zero), and a q x 1 vector of parameters φ (not expected to be zero). Note that q could be

zero (i.e. the set of parameters not expected to be zero may be empty).

In one embodiment, the input data is organised into an $n \times p$ data matrix $X=(x_{ij})$ with n test training samples and p components. Typically, p will be much greater than n.

In another embodiment, data matrix X may be replaced by an n x n kernel matrix K to obtain smooth functions of X as predictors instead of linear predictors. An example of the kernel matrix K is $k_{ij}=\exp(-0.5*(x_i-x_j)^t(x_i-x_j)/\sigma^2)$ where the subscript on x refers to a row number in the matrix X. Ideally, subsets of the columns of K are selected which give sparse representations of these smooth functions.

Typically, as discussed above, the component weights are estimated in a manner which takes into account the apriori assumption that most of the component weights are zero.

In one embodiment, the prior specified for the component weights is of the form:

$$p(\beta)=\int_{v^2} p(\beta|v^2)p(v^2)dv^2 \qquad \text{(C1)}$$

wherein v is a p x 1 vector of hyperparameters, and where $p(\beta|v^2)$ is $N(0,\mathrm{diag}\{v^2\})$ and $p(v^2)$ is some hyperprior distribution for v^2.

A suitable form of hyperprior is, $p(v^2)\alpha\prod_{i=1}^{p}p(v_i^2)$. Jeffreys

In another embodiment, the hyperprior $p(v^2)$ is such that each $t_i^2=1/v_i^2$ has an independent gamma distribution.

In another embodiment, the hyperprior $p(v^2)$ is such that each v_i^2 has an independent gamma distribution.

Preferably, an uninformative prior for φ is specified.

The likelihood function is defined from a model for the distribution of the data. Preferably, in general, the likelihood function is any suitable likelihood function. For example, the likelihood function $L(\underset{\sim}{y}|\beta\varphi)$ may be, but not restricted to, of the form appropriate for a generalised linear model (GLM), such as for example, that described by Nelder and Wedderburn (1972). Preferably, in this case, the likelihood function is of the form:

$$L = \log p(y \mid \beta, \phi) = \sum_{i=1}^{N} \{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \} \qquad \text{(C2)}$$

where $y = (y_1, \ldots, y_n)^T$ and $a_i(\phi) = \phi / w_i$ with the w_i being a fixed set of known weights and ϕ a single scale parameter.

Preferably, the likelihood function is specified as follows: We have

$$E\{y_i\} = b'(\theta_i)$$
$$Var\{y\} = b''(\theta_i) a_i(\phi) = \tau_i^2 a_i(\phi) \qquad \text{(C3)}$$

Each observation has a set of covariates x_i and a linear predictor $\eta_i = x_i^T \beta$. The relationship between the mean of the i^{th} observation and its linear predictor is given by the link function $\eta_i = g(\mu_i) = g(b'(\theta_i))$. The inverse of the link is denoted by h, i.e
$\mu_i = b'(\theta_i) = h(\eta_i)$.

In addition to the scale parameter, a generalised linear model may be specified by four components:

- the likelihood or (scaled) deviance function,
- the link function
- the derivative of the link function
- the variance function.

Some common examples of generalised linear models are given in the table below.

Distribution	Link function $g(\mu)$	Derivative of link function	Variance function	Scale parameter
Gaussian	μ	1	1	Yes
Binomial	$\log(\mu/(1-\mu))$	$1/(\mu(1-\mu))$	$\mu(1-\mu)/n$	No
Poisson	$\log(\mu)$	$1/\mu$	μ	No
Gamma	$1/\mu$	$-1/\mu^2$	μ^2	Yes
Inverse Gaussian	$1/\mu^2$	$-2/\mu^3$	μ^3	Yes

In another embodiment, a quasi likelihood model is specified wherein only the link function and variance function are defined. In some instances, such specification results in the models in the table above. In other instances, no distribution is specified.

In one embodiment, the posterior distribution of β φ and ν given y is estimated using:

$$p(\beta\varphi\nu|y) \propto L(y|\beta\varphi)p(\beta|\nu)p(\nu) \qquad \text{(C4)}$$

wherein $L(\underset{\sim}{y}|\beta\varphi)$ is the likelihood function.

In one embodiment, ν may be treated as a vector of missing data and an iterative procedure used to maximise equation (C4) to produce maximum a posteriori estimates of β. The prior of equation (C1) is such that the maximum a posteriori estimates will tend to be sparse i.e. if a large number of parameters are redundant, many components of β will be zero.

As stated above, the component weights which maximise the posterior distribution may be determined using an iterative procedure. Preferable, the iterative procedure for

maximising the posterior distribution of the components and component weights is an EM algorithm comprising an E step and an M step, such as, for example, that described in Dempster et al, 1977.

In conducting the EM algorithm, the E step preferably comprises the step of computing terms of the form

$$P = \sum_{i=1}^{p} E\{\beta_i^2/\nu_i^2 \mid \hat{\beta}_i\} = \sum_{i=1}^{p} \beta_i^2/\hat{d}_i^2 \qquad \text{(C4a)}$$

where $\hat{d}_i = d_i(\hat{\beta}_i) = E\{1/\nu_i^2 \mid \hat{\beta}_i\}^{-0.5}$ and for convenience we define $\hat{d}_i = 1/\hat{d}_i = 0$ if $\hat{\beta}_i = 0$. In the following we write $\hat{d} = d(\hat{\beta}) = (\hat{d}_1, \hat{d}_2, ..., \hat{d}_n)^T$. In a similar manner we define for example $d(\beta^{(n)})$ and $d(\gamma^{(n)}) = P_n^T d(P_n \gamma^{(n)})$ where $\beta^{(n)} = P_n \gamma^{(n)}$ and P_n is obtained from the p by p identity matrix by omitting columns j for which $\beta_j^{(n)} = 0$.

Preferably, equation (C4a) is computed by calculating the conditional expected value of $t_i^2 = 1/\nu_i^2$ when $p\left(\beta_i \middle| \nu_i^2\right)$ is $N\left(0, \nu_i^2\right)$ and $p\left(\nu_i^2\right)$ has a specified prior distribution. Specific examples and formulae will be given later.

In a general embodiment, appropriate for any suitable likelihood function, the EM algorithm comprises the steps:

(a) Selecting a hyperprior and values for its parameters. Initialising the algorithm by setting n=0, $S_0 = \{1, 2, ..., p\}$, initialise $\varphi^{(0)}$, β^* and applying a value for ε, such as for example $\varepsilon = 10^{-5}$;

(b) Defining

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases} \qquad \text{(C5)}$$

and let Pn be a matrix of zeroes and ones such that the nonzero elements $\gamma^{(n)}$ of $\beta^{(n)}$ satisfy

$$\gamma^{(n)} = P_n^T \beta^{(n)} \quad , \quad \beta^{(n)} = P_n \gamma^{(n)}$$
$$\gamma = P_n^T \beta \quad , \quad \beta = P_n \gamma$$

(c) performing an estimation (E) step by calculating the conditional expected value of the posterior distribution of component weights using the function:

$$Q(\beta \mid \beta^{(n)}, \varphi^{(n)}) = E\{ \log p(\beta, \varphi, \nu \mid y) \mid y, \beta^{(n)}, \phi^{(n)}\} = L(y \mid \beta, \phi^{(n)}) - 0.5\, \beta^T \Delta(d(\beta^{(n)}))^{-2} \beta \qquad \text{(C6)}$$

where L is the log likelihood function of y. Using $\beta = P_n \gamma$ and $d(\gamma^{(n)})$ as defined in (C4a), (C6) can be written as

$$Q(\gamma \mid \gamma^{(n)}, \phi^{(n)}) = L(y \mid P_n \gamma, \phi^{(n)}) - 0.5 \gamma^T \Delta(d(\gamma^{(n)}))^{-2} \gamma \qquad \text{(C7)}$$

(d) performing a maximisation (M) step by applying an iterative procedure to maximise Q as a function of γ whereby $\gamma_0 = \gamma^{(n)}$ and for r=0,1,2, $\gamma_{r+1} = \gamma_r + \alpha_r \delta r$ and where αr is chosen by a line search algorithm to ensure

$$Q(\gamma_{r+1} \mid \gamma^{(n)}, \phi^{(n)}) > Q(\gamma_r \mid \gamma^{(n)}, \phi^{(n)}), \text{ and}$$

$$\delta_r = \Delta(d(\gamma^{(n)}))[-\Delta(d(\gamma^{(n)}))\frac{\partial^2 L}{\partial^2 \gamma_r}\Delta(d(\gamma^{(n)})) + I]^{-1}(\Delta(d(\gamma^{(n)}))(\frac{\partial L}{\partial \gamma_r} - \frac{\gamma_r}{d(\gamma^{(n)})}) \qquad \text{(C8)}$$

where:

$d(\gamma^{(n)}) = P_n^T d(P_n \gamma^{(n)})$ as in (C4a); and

$$\frac{\partial L}{\partial \gamma_r} = P_n^T \frac{\partial L}{\partial \beta_r} \quad , \quad \frac{\partial^2 L}{\partial^2 \gamma_r} = P_n^T \frac{\partial^2 L}{\partial^2 \beta_r} P_n$$

for $\beta_r = P_n \gamma_r$.

(e) Let γ^* be the value of γr when some convergence criterion is satisfied, for example, $||\gamma r - \gamma r+1|| < \varepsilon$ (for example 10^{-5});

(f) Defining $\beta^* = P_n\gamma^*$, $S_{n+1}=\{i: |\beta_i| > \max_j(|\beta_j|*\epsilon_1)\}$

where ε_1 is a small constant, for example 1e-5.

(g) Set n=n+1 and choose $\varphi^{(n+1)} = \varphi^{(n)}+\kappa_n(\varphi^* - \varphi^{(n)})$

where φ^* satisfies $\frac{\partial}{\partial\phi}L(y|P_n\gamma^*,\phi)=0$ and κ_n is a damping factor such that $0< \kappa_n \leq 1$; and

(h) Check convergence. If $||\gamma^* - \gamma^{(n)}|| < \varepsilon_2$ where ε_2 is suitably small then stop, else go to step (b) above.

In another embodiment, $t_i^2 = 1/\nu_i^2$ has an independent *gamma distribution* with scale parameter b>0 and shape parameter k>0 so that the density of t_i^2 is

$$\gamma(t_i^2,b,k)=b^{-1}(t_i^2/b)^{k-1}\exp(-t_i^2/b)/\Gamma(k)$$

It can be shown that

$$E\{t^2|\beta\} = (2k+1)/(2/b+\beta^2)$$

as follows:

Define

$$I(p,b,k)=\int_0^\infty (t^2)^p\, t\exp(-0.5\beta^2t^2)\gamma(t^2,b,k)dt^2$$

then

$$I(p,b,k)=b^{p+0.5}\{\Gamma(p+k+0.5)/\Gamma(k)\}(1+0.5b\beta^2)^{-(p+k+0.5)}$$

Proof

Let $s = \beta^2/2$ then

$$I(p,b,k)=b^{p+0.5}\int_0^{\infty}(t^2/b)^{p+0.5}\exp(-st^2)\gamma(t^2,b,k)dt^2$$

Now using the substitution $u = t^2/b$ we get

$$I(p,b,k)=b^{p+0.5}\int_0^{\infty}(u)^{p+0.5}\exp(-sub)\gamma(u,1,k)du$$

Now let $s'=bs$ and substitute the expression for $\gamma(u,1,k)$. This gives

$$I(p,b,k)=b^{p+0.5}\int_0^{\infty}\exp(-(s'+1)u)u^{p+k+0.5-1}du / \Gamma(k)$$

Looking up a table of Laplace transforms, eg Abramowitz and Stegun, then gives the result.

The conditional expectation follows from

$$E\{ t^2 \mid \beta \} = I(1,b,k)/I(0,b,k)$$
$$= (2k+1)/(2/b + \beta^2)$$

As k tends to zero and b tends to infinity we get the equivalent result using Jeffreys prior. For example, for $k=0.005$ and $b=2*10^5$

$$E\{ t^2 \mid \beta \} = (1.01)/(10^{-5} + \beta^2)$$

Hence we can get arbitrarily close to the algorithm with a Jeffreys hyperprior with this proper prior.

In another embodiment, v_i^2 has an independent *gamma distribution* with scale parameter $b>0$ and shape parameter $k>0$. It can be shown that

$$E\{\nu_i^{-2}|\beta_i\}=\frac{\int_0^\infty u^{k-3/2-1}\exp(-(\lambda_i/u+u))\,du}{b\int_0^\infty u^{k-1/2-1}\exp(-(\lambda_i/u+u))\,du}$$
$$=\sqrt{\frac{2}{b}}\frac{1}{|\beta_i|}\frac{K_{3/2-k}(2\sqrt{\lambda_i})}{K_{1/2-k}(2\sqrt{\lambda_i})} \qquad \text{(C9)}$$
$$=\frac{1}{|\beta_i|^2}\frac{(2\sqrt{\lambda_i})K_{3/2-k}(2\sqrt{\lambda_i})}{K_{1/2-k}(2\sqrt{\lambda_i})}$$

where $\lambda_i=\beta_i^2/2b$ and K denotes a modified Bessel function, which can be shown as follows:

For k=1 in equation (c9)

$$E\{\nu_i^{-2}|\beta_i\}=\sqrt{2/b}(1/|\beta_i|)$$

For K=0.5 in equation (C9)

$$E\{\nu_i^{-2}|\beta_i\}=\sqrt{2/b}(1/|\beta_i|)\{K_1(2\sqrt{\lambda_i})/K_0(2\sqrt{\lambda_i})\}$$

or equivalently

$$E\{\nu_i^{-2}|\beta_i\}=(1/|\beta_i|^2)\{2\sqrt{\lambda_i}K_1(2\sqrt{\lambda_i})/K_0(2\sqrt{\lambda_i})\}$$

Proof

From the definition of the conditional expectation, writing $\lambda_i=\beta_i^2/2b$, we get

$$E\{\nu_i^{-2}|\beta_i\}=\frac{\int_0^\infty \nu_i^{-2}\nu_i^{-1}\exp(-\lambda_i\nu_i^{-2})b^{-1}(\nu_i^{-2}/b)^{k-1}\exp(\nu_i^{-2}/b)d\nu_i^{2}}{\int_0^\infty \nu_i^{-1}\exp(-\lambda_i\nu_i^{-2})b^{-1}(\nu_i^{-2}/b)^{k-1}\exp(\nu_i^{-2}/b)d\nu_i^{2}}$$

Rearranging, simplifying and making the substitution $u=\nu_i^2/b$ gives A.1

The integrals in A.1 can be evaluated by using the result

$$\int_0^\infty x^{-b-1}\exp\left[-\left(x+\frac{a^2}{x}\right)\right]dx = \frac{2}{a^b}K_b(2a)$$

where K denotes a modified Bessel function, see Watson(1966).

Examples of members of this class are k=1 in which case

$$E\{\nu_i^{-2}|\beta_i\} = \sqrt{2/b}(1/|\beta_i|)$$

which corresponds to the prior used in the Lasso technique, Tibshirani(1996). See also Figueiredo(2001).

The case k=0.5 gives

$$E\{\nu_i^{-2}|\beta_i\} = \sqrt{2/b}(1/|\beta_i|)\{K_1(2\sqrt{\lambda_i})/K_0(2\sqrt{\lambda_i})\}$$

or equivalently

$$E\{\nu_i^{-2}|\beta_i\} = (1/|\beta_i|^2)\{2\sqrt{\lambda_i}K_1(2\sqrt{\lambda_i})/K_0(2\sqrt{\lambda_i})\}$$

where K_0 and K_1 are modified Bessel functions, see Abramowitz and Stegun(1970). Polynomial approximations for evaluating these Bessel functions can be found in Abramowitz and Stegun(1970, p379). Details of the above calculations are given in the Appendix.

The expressions above demonstrate the connection with the Lasso model and the Jeffreys prior model.

It will be appreciated by those skilled in the art that as k tends to zero and b tends to infinity the prior tends to a Jeffreys improper prior.

In one embodiment, the priors with $0< k \leq 1$ and $b>0$ form a class of priors which might be interpreted as penalising non zero coefficients in a manner which is between the Lasso prior and the original specification using Jeffreys prior.

In another embodiment, in the case of generalised linear models, step (d) in the maximisation step may be estimated by replacing $\frac{\partial^2 L}{\partial^2 \gamma_r}$ with its expectation $E\{\frac{\partial^2 L}{\partial^2 \gamma_r}\}$. This is preferred when the model of the data is a generalised linear model.

For generalised linear models the expected value $E\{\frac{\partial^2 L}{\partial^2 \gamma_r}\}$ may be calculated as follows. Beginning with

$$\frac{\partial L}{\partial \beta} = X^T\{\Delta(\frac{1}{\tau_i^2}\frac{\partial \mu_i}{\partial \eta_i})(\frac{y_i - \mu_i}{a_i(\phi)})\} \qquad \text{(C10)}$$

where X is the N by p matrix with i^{th} row x_i^T and

$$E\{\frac{\partial^2 L}{\partial^2 \beta^2}\} = -E\{(\frac{\partial L}{\partial \beta})(\frac{\partial L}{\partial \beta})^T\} \qquad \text{(C11)}$$

we get

$$E\{\frac{\partial^2 L}{\partial^2 \beta^2}\} = -X^T\Delta(a_i(\varphi)\tau_i^2(\frac{\partial \eta_i}{\partial \mu_i})^2)^{-1}X$$

Equations (C10) and (C11) can be written as

$$\frac{\partial L}{\partial \beta} = X^t V^{-1}(\frac{\partial \eta}{\partial \mu})(y-\mu) \qquad \text{(C12)}$$

$$E\{\frac{\partial^2 L}{\partial \beta^2}\} = -X^t V^{-1} X \qquad \text{(C13)}$$

where $V=\Delta(a_i(\varphi)\tau_i^2(\frac{\partial \eta_i}{\partial \mu_i})^2)$.

Preferably, for generalised linear models, the EM algorithm comprises the steps:

(a) Choose a hyper prior and its parameters. Initialising the algorithm by setting n=0, $S_0 = \{1,2,\ldots, p\}$, $\varphi^{(0)}$, applying a value for ε, such as for example $\varepsilon = 10^{-5}$, and

If $p \leq N$ compute initial values β^* by

$$\beta^* = (X^t X + \lambda I)^{-1} X^T g(y+\zeta) \qquad \text{(C14)}$$

and if $p > N$ compute initial values β^* by

$$\beta^* = \frac{1}{\lambda}(I - X^T(XX^T + \lambda I)^{-1}X)X^T g(y+\zeta) \qquad \text{(C15)}$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$ and ζ is small and chosen so that the link function is well defined at $y+\zeta$.

(b) Defining

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

and let Pn be a matrix of zeroes and ones such that the nonzero elements $\gamma(n)$ of $\beta(n)$ satisfy

$$\gamma^{(n)} = P_n^T \beta^{(n)} \quad , \quad \beta^{(n)} = P_n \gamma^{(n)}$$
$$\gamma = P_n^T \beta \quad , \quad \beta = P_n \gamma$$

(c) performing an estimation (E) step by calculating the conditional expected value of the posterior distribution of component weights using the function:

$$\begin{aligned} Q(\beta \mid \beta^{(n)}, \varphi^{(n)}) &= E\{ \log p(\beta, \varphi, \nu \mid y) \mid y, \beta^{(n)}, \phi^{(n)}\} \\ &= L(y \mid \beta, \phi^{(n)}) - 0.5\, \beta^T \Delta(d(\beta^{(n)}))^{-2} \beta \end{aligned} \qquad \text{(C16)}$$

where L is the log likelihood function of y. Using $\beta = P_n\gamma$ and $\beta^{(n)} = P_n\gamma^{(n)}$ (C16) can be written as

$$Q(\gamma \mid \gamma^{(n)}, \phi^{(n)}) = L(y \mid P_n\gamma, \phi^{(n)}) - 0.5\, \gamma^T \Delta(d(\gamma^{(n)}))^{-2}\gamma \qquad \text{(C17)}$$

(d) performing a maximisation (M) step by applying an iterative procedure, for example a Newton Raphson iteration, to maximise Q as a function of γ whereby $\gamma_0 = \gamma^{(n)}$ and for r=0,1,2,... $\gamma_{r+1} = \gamma_r + \alpha_r \delta_r$ where α_r is chosen by a line search algorithm to ensure $Q(\gamma_{r+1} \mid \gamma^{(n)}, \phi^{(n)}) > Q(\gamma_r \mid \gamma^{(n)}, \phi^{(n)})$, and for $p \leq N$ use

$$\delta_r = \Delta(d(\gamma^{(n)}))[Y_n^T V_r^{-1} Y_n + I]^{-1}(Y_n^T V_r^{-1} z_r - \frac{\gamma_r}{d(\gamma^{(n)})}) \qquad \text{(C18)}$$

where

$$Y_n = \Delta(d(\gamma^{(n)}))P_n^T X$$

$$V = \Delta(a_i(\varphi)\tau_i^2(\frac{\partial \eta_i}{\partial \mu_i})^2)$$

$$z = \frac{\partial \eta}{\partial \mu}(y-\mu)$$

and the subscript r denotes that these quantities are evaluated at $\mu = h(XP_n\gamma_r)$.

For $p > N$ use

$$\delta_r = \Delta(d(\gamma^{(n)}))[I - Y_n^T(Y_n Y_n^T + V_r)^{-1}Y_n](Y_n^T V_r^{-1} z_r - \frac{\gamma_r}{d(\gamma^{(n)})}) \qquad \text{(C19)}$$

with V_r and z_r defined as before.

(e) Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g

$$||\gamma_r - \gamma_{r+1}|| < \varepsilon \quad \text{(for example } 10^{-5}\text{)}.$$

(f) Define $\beta^* = P_n\gamma^*$, $S_{n+1}=\{i: |\beta_i| > \max_j(|\beta_j|*\epsilon_1)\}$ where ϵ_1 is a small constant, say 1e-5. Set n=n+1 and choose $\varphi^{n+1} = \varphi^n + \kappa_n(\varphi^* - \varphi^n)$ where ϕ^* satisfies $\frac{\partial}{\partial\phi}L(y|P_n\gamma^*,\phi)=0$ and κ_n is a damping factor such that $0< \kappa_n \leq 1$. Note that in some cases the scale parameter is known or this equation can be solve explicitly to get an updating equation for φ.

The above embodiments may be extended to incorporate quasi likelihood methods Wedderburn (1974) and McCullagh and Nelder (1983)). In such an embodiment, the same iterative procedure as detailed above will be appropriate, but with L replaced by a quasi likelihood as shown above and, for example, Table 8.1 in McCullagh and Nelder (1983). In one embodiment there is a modified updating method for the scale parameter φ. To define these models requires specification of the variance function τ^2 , the link function g and the derivative of the link function $\frac{\partial\eta}{\partial\mu}$. Once these are defined the above algorithm can be applied.

In one embodiment for quasi likelihood models, step 5 of the above algorithm is modified so that the scale parameter is updated by calculating

$$\phi^{(n+1)}=\frac{1}{N}\sum_{i=1}^{N}\frac{(y_i-\mu_i)^2}{\tau_i^2}$$

where μ and τ are evaluated at $\beta^* = P_n\gamma^*$. Preferably, this updating is performed when the number of parameters s in the model is less than N. A divisor of N -s can be used when s is much less than N.

In another embodiment, for both generalised linear models and Quasi likelihood models the covariate matrix X with rows x_i^T can be replaced by a matrix K with ij^{th} element k_{ij} and $k_{ij} = \kappa(x_i - x_j)$ for some kernel function κ . This matrix can also be augmented with a vector of ones. Some example kernels are given in Table 2 below, see Evgeniou et al(1999).

Kernel function	Formula for $\kappa(x - y)$
Gaussian radial basis function	$\exp(- \|\| x - y \|\|^2 / a)$, $a>0$
Inverse multiquadric	$(\|\| x - y \|\|^2 + c^2)^{-1/2}$
Multiquadric	$(\|\| x - y \|\|^2 + c^2)^{1/2}$
Thin plate splines	$\|\| x - y \|\|^{2n+1}$ $\|\| x - y \|\|^{2n} \ln(\|\| x - y \|\|)$
Multi layer perceptron	$\tanh(x \cdot y - \theta)$, for suitable θ
Ploynomial of degree d	$(1 + x \cdot y)^d$
B splines	$B_{2n+1}(x - y)$
Trigonometric polynomials	$\sin((d +1/2)(x-y))/\sin((x-y)/2)$

Table 2: Examples of kernel functions

In Table 2 the last two kernels are one dimensional i.e. for the case when X has only one column. Multivariate versions can be derived from products of these kernel functions. The definition of B_{2n+1} can be found in De Boor(1978). Use of a kernel function in either a generalised linear model or a quasi likelihood model results in mean values which are smooth (as opposed to transforms of linear) functions of the covariates X. Such models may give a substantially better fit to the data.

A fourth embodiment relating to a proportional hazards model will now be described.

D. Proportional Hazard Models

The method of this embodiment may utilise training samples in order to identify a subset of components which are capable of affecting the probability that a defined event (eg death, recovery) will occur within a certain time period. Training samples are obtained from a system and the time measured from when the training sample is obtained to when the event has occurred. Using a statistical method to associate the time to the event with the data obtained from a plurality of training samples, a subset of components may be identified that are capable of predicting the distribution of the time to the event. Subsequently, knowledge of the subset of components can be used for tests, for example clinical tests to predict for example, statistical features of the time to death or time to relapse of a disease. For example, the data from a subset of components of a system may be obtained from a DNA microarray. This data may be used to predict a clinically relevant event such as, for example, expected or median patient survival times, or to predict onset of certain symptoms, or relapse of a disease.

In this way, the present invention identifies preferably a relatively small number of components which can be used to predict the distribution of the time to an event of a system. The selected components are "predictive" for that time to an event. Essentially, from all the data which is generated from the system, the method of the present invention enables identification of a small number of components which can be used to predict time to an event. Once those components have been identified by this method, the components can be used in future to predict statistical

features of the time to an event of a system from new samples. The method of the present invention preferably utilises a statistical method to eliminate components that are not required to correctly predict the time to an event of a system. By appropriate selection of the hyperparameters in the model some control over the size of the selected subset can be achieved.

As used herein, "time to an event" refers to a measure of the time from obtaining the sample to which the method of the invention is applied to the time of an event. An event may be any observable event. When the system is a biological system, the event may be, for example, time till failure of a system, time till death, onset of a particular symptom or symptoms, onset or relapse of a condition or disease, change in phenotype or genotype, change in biochemistry, change in morphology of an organism or tissue, change in behaviour.

The samples are associated with a particular time to an event from previous times to an event. The times to an event may be times determined from data obtained from, for example, patients in which the time from sampling to death is known, or in other words, "genuine" survival times, and patients in which the only information is that the patients were alive when samples were last obtained, or in other words, "censored" survival times indicating that the particular patient has survived for at least a given number of days.

In one embodiment, the input data is organised into an $n \times p$ data matrix $X = (x_{ij})$ with n test training samples and p components. Typically, p will be much greater than n.

For example, consider an $N \times p$ data matrix $X = (x_{ij})$ from, for example, a microarray experiment, with N individuals (or samples) and the same p genes for each individual.

Preferably, there is associated with each individual $i\ (i=1,2,\cdots,N)$ a variable $y_i\ (y_i \geq 0)$ denoting the time to an event, for example, survival time. For each individual there may also be defined a variable that indicates whether that individual's survival time is a genuine survival time or a censored survival time. Denote the censor indicators as c_i where

$$c_i = \begin{cases} 1, \text{if } y_i \text{ is uncensored} \\ 0, \text{if } y_i \text{ is censored} \end{cases}$$

The Nx1 vector with survival times y_i may be written as $\underset{\sim}{y}$ and the Nx1 vector with censor indicators c_i as $\underset{\sim}{c}$.

Typically, as discussed above, the component weights are estimated in a manner which takes into account the a priori assumption that most of the component weights are zero.

Preferably, the prior specified for the component weights is of the form

$$P(\beta_1,\beta_2,\cdots,\beta_n) = \int_{\tau}\prod_{i=1}^{N} P(\beta_i|\tau_i)P(\tau_i)d\tau \qquad \text{(D1)}$$

where $\beta_1,\beta_2,\cdots,\beta_n$ are component weights, $P(\beta_i|\tau_i)$ is $N(0,\tau_i^2)$ and $P(\tau_i)$ is some hyperprior distribution

$$P(\tau) = \prod_{i=1}^{n} P(\tau_i)$$

that is not a Jeffrey's hyperprior.

In one embodiment, the prior distribution is an inverse gamma prior for τ in which $t_i^2 = 1/\tau_i^2$ has an independent *gamma distribution* with scale parameter b>0 and shape parameter k>0 so that the density of t_i^2 is

$$\gamma(t_i^2,b,k)=b^{-1}(t_i^2/b)^{k-1}\exp(-t_i^2/b)/\Gamma(k)\ .$$

It can be shown that:

$$E\{\ t^2\ |\ \beta\ \} = \ (2k+1)/(2/b+\beta^2) \qquad \text{(A)}$$

Equation A can be shown as follows:

Define

$$I(p,b,k)=\int_0^\infty (t^2)^p\ t\ \exp(-0.5\beta^2t^2)\gamma(t^2,b,k)dt^2$$

then

$$I(p,b,k)=b^{p+0.5}\{\Gamma(p+k+0.5)/\Gamma(k)\}(1+0.5b\beta^2)^{-(p+k+0.5)}$$

Proof

Let $s=\beta^2/2$ then

$$I(p,b,k)=b^{p+0.5}\int_0^\infty (t^2/b)^{p+0.5}\exp(-st^2)\gamma(t^2,b,k)dt^2$$

Now using the substitution $u = t^2/b$ we get

$$I(p,b,k)=b^{p+0.5}\int_0^\infty (u)^{p+0.5}\exp(-sub)\gamma(u,1,k)du$$

Now let $s'=bs$ and substitute the expression for $\gamma(u,1,k)$. This gives

$$I(p,b,k)=b^{p+0.5}\int_0^\infty \exp(-(s'+1)u)u^{p+k+0.5-1}du\ /\ \Gamma(k)$$

Looking up a table of Laplace transforms, eg Abramowitz and Stegun, then gives the result.

The conditional expectation follows from

$$E\{\ t^2\ |\ \beta\ \} = I(1,b,k)/I(0,b,k)$$
$$= (2k+1)/(2/b+\beta^2)$$

As k tends to zero and b tends to infinity, a result equivalent to using Jeffreys prior is obtained. For example, for k=0.005 and b=2*10^5

$$E\{ t^2 | \beta \} = (1.01)/(10^{-5} + \beta^2)$$

Hence we can get arbitrarily close to a Jeffery's prior with this proper prior.

The modified algorithm for this model has

$$b^{(n)} = E\{t^2 | \beta^{(n)}\}^{-0.5}$$

where the expectation is calculated as above.

In yet another embodiment, the prior distribution is a gamma distribution for τ_{ig}^2. Preferably, the gamma distribution has scale parameter b>0 and shape parameter k>0.

It can be shown, that

$$E\{\tau_i^{-2} | \beta_i\} = \frac{\int_0^\infty u^{k-3/2-1} \exp(-(\gamma_i/u + u))\, du}{b \int_0^\infty u^{k-1/2-1} \exp(-(\gamma_i/u + u))\, du}$$

$$= \sqrt{\frac{2}{b}} \frac{1}{|\beta_i|} \frac{K_{3/2-k}(2\sqrt{\gamma_i})}{K_{1/2-k}(2\sqrt{\gamma_i})}$$

$$= \frac{1}{|\beta_i|^2} \frac{(2\sqrt{\gamma_i}) K_{3/2-k}(2\sqrt{\gamma_i})}{K_{1/2-k}(2\sqrt{\gamma_i})}$$

where $\gamma_i = \beta_i^2/2b$ and K denotes a modified Bessel function. Some special members of this class are k=1 in which case

$$E\{\tau_i^{-2} | \beta_i\} = \sqrt{2/b}\,(1/|\beta_i|)$$

which corresponds to the prior used in the Lasso technique, Tibshirani(1996). See also Figueiredo(2001).

The case k=0.5 gives

$$E\{\tau_i^{-2}|\beta_i\} = \sqrt{2/b}(1/|\beta_i|)\{K_1(2\sqrt{\gamma_i})/K_0(2\sqrt{\gamma_i})\}$$

or equivalently

$$E\{\tau_i^{-2}|\beta_i\} = (1/|\beta_i|^2)\{2\sqrt{\gamma_i}K_1(2\sqrt{\gamma_i})/K_0(2\sqrt{\gamma_i})\}$$

where K_0 and K_1 are modified Bessel functions, see Abramowitz and Stegun(1970). Polynomial approximations for evaluating these Bessel functions can be found in Abramowitz and Stegun(1970, p379).

The expressions above demonstrate the connection with the Lasso model and the Jeffreys prior model.

Details of the above calculations are as follows:

For the gamma prior above and $\gamma_i = \beta_i^2/2b$

$$\begin{aligned} E\{\tau_i^{-2}|\beta_i\} &= \frac{\int_0^\infty u^{k-3/2-1}\exp(-(\gamma_i/u+u))\,du}{b\int_0^\infty u^{k-1/2-1}\exp(-(\gamma_i/u+u))\,du} \\ &= \sqrt{\frac{2}{b}}\frac{1}{|\beta_i|}\frac{K_{3/2-k}(2\sqrt{\gamma_i})}{K_{1/2-k}(2\sqrt{\gamma_i})} \\ &= \frac{1}{|\beta_i|^2}\frac{(2\sqrt{\gamma_i})K_{3/2-k}(2\sqrt{\gamma_i})}{K_{1/2-k}(2\sqrt{\gamma_i})} \end{aligned} \qquad \text{(D2)}$$

where K denotes a modified Bessel function.
For k=1 in (D2)

$$E\{\tau_i^{-2}|\beta_i\} = \sqrt{2/b}(1/|\beta_i|)$$

For K=0.5 in (D2)

$$E\{\tau_i^{-2}|\beta_i\} = \sqrt{2/b}(1/|\beta_i|)\{K_1(2\sqrt{\gamma_i})/K_0(2\sqrt{\gamma_i})\}$$

or equivalently

$$E\{\tau_i^{-2}|\beta_i\} = (1/|\beta_i|^2)\{2\sqrt{\gamma_i}K_1(2\sqrt{\gamma_i})/K_0(2\sqrt{\gamma_i})\}$$

Proof

From the definition of the conditional expectation, writing $\gamma_i=\beta_i^2/2b$, we get

$$E\{\tau_i^{-2}|\beta_i\} = \frac{\int_0^\infty \tau_i^{-2}\tau_i^{-1}\exp(-\gamma_i\tau_i^{-2})b^{-1}(\tau_i^{-2}/b)^{k-1}\exp(\tau_i^{-2}/b)d\tau_i^2}{\int_0^\infty \tau_i^{-1}\exp(-\gamma_i\tau_i^{-2})b^{-1}(\tau_i^{-2}/b)^{k-1}\exp(\tau_i^{-2}/b)d\tau_i^2}$$

Rearranging, simplifying and making the substitution $u=\nu_i^2/b$ gives A.1

The integrals in A.1 can be evaluated by using the result

$$\int_0^\infty x^{-b-1}\exp\left[-\left(x+\frac{a^2}{x}\right)\right]dx = \frac{2}{a^b}K_b(2a)$$

where K denotes a modified Bessel function, see Watson(1966).

In a particularly preferred embodiment, $p(\tau_i)\alpha 1/\tau_i^2$ is a Jeffreys prior, Kotz and Johnson(1983).

The likelihood function defines a model which fits the data based on the distribution of the data. Preferably, the likelihood function is of the form:

$$Log\ (Partial)\ Likelihood = \sum_{i=1}^{N} g_i\left(\underset{\sim}{\beta},\underset{\sim}{\varphi};X,\underset{\sim}{y},\underset{\sim}{c}\right)$$

where $\underset{\sim}{\beta}^T = (\beta_1, \beta_2, \cdots, \beta_p)$ and $\underset{\sim}{\varphi}^T = (\varphi_1, \varphi_2, \cdots, \varphi_q)$ are the model parameters. The model defined by the likelihood function may be any model for predicting the time to an event of a system.

In one embodiment, the model defined by the likelihood is Cox's proportional hazards model. Cox's proportional hazards model was introduced by Cox (1972) and may preferably be used as a regression model for survival data. In Cox's proportional hazards model, $\underset{\sim}{\beta}'$ is a vector of (explanatory) parameters associated with the components. Preferably, the method of the present invention provides for the parsimonious selection (and estimation) from the parameters $\underset{\sim}{\beta}' = (\beta_1, \beta_2, \cdots, \beta_p)$ for Cox's proportional hazards model given the data X, $\underset{\sim}{y}$ and $\underset{\sim}{c}$.

Application of Cox's proportional hazards model can be problematic in the circumstance where different data is obtained from a system for the same survival times, or in other words, tied survival times. Tied survival times may be subjected to a pre-processing step that leads to unique survival times. The pre-processing proposed simplifies the ensuing code as it avoids concerns about tied survival times in the subsequent application of Cox's proportional hazards model.

The pre-processing of the survival times applies by adding an extremely small amount of insignificant random noise. Preferably, the procedure is to take sets of tied times and add to each tied time within a set of tied times a random amount that is drawn from a normal distribution that has zero mean and variance proportional to the smallest non-zero distance between sorted survival times. Such pre-processing achieves an elimination of tied times without imposing a draconian perturbation of the survival times.

The pre-processing generates distinct survival times. Preferably, these times may be ordered in increasing magnitude denoted as $\underset{\sim}{t}=\left(t_{(1)},t_{(2)},\cdots t_{(N)}\right),\ t_{(i+1)}>t_{(i)}$.

Denote by Z the Nxp matrix that is the re-arrangement of the rows of X where the ordering of the rows of Z corresponds to the ordering induced by the ordering of $\underset{\sim}{t}$; also denote by Z_j the j^{th} row of the matrix Z. Let d be the result of ordering c with the same permutation required to order $\underset{\sim}{t}$.

After pre-processing for tied survival times is taken into account and reference is made to standard texts on survival data analysis (eg Cox and Oakes, 1984), the likelihood function for the proportional hazards model may preferably be written as

$$l\left(\underset{\sim}{t}\,|\,\underset{\sim}{\beta}\right)=\prod_{j=1}^{N}\left(\frac{exp\left(Z_j\underset{\sim}{\beta}\right)}{\sum_{i\in\Re_j}exp\left(Z_i\underset{\sim}{\beta}\right)}\right)^{d_j} \qquad \text{(D3)}$$

where $\underset{\sim}{\beta}^T=(\beta_1,\beta_2,\cdots,\beta_n)$, Z_j = the j^{th} row of Z, and $\Re_j=\{i:i=j,j+1,\cdots,N\}$ = the risk set at the j^{th} ordered event time $t_{(j)}$.

The logarithm of the likelihood (ie $L=log(l)$) may preferably be written as

$$L\left(\underset{\sim}{t}\,|\,\underset{\sim}{\beta}\right)=\sum_{i=1}^{N}d_i\left(Z_i\underset{\sim}{\beta}-log\left(\sum_{j\in\Re_i}exp\left(Z_j\underset{\sim}{\beta}\right)\right)\right)$$

$$=\sum_{i=1}^{N}d_i\left(Z_i\underset{\sim}{\beta}-log\left(\sum_{j=1}^{N}\zeta_{i,j}\,exp\left(Z_j\underset{\sim}{\beta}\right)\right)\right), \qquad \text{(D4)}$$

where

$$\zeta_{i,j}=\begin{cases}0, & \text{if } j<i\\ 1, & \text{if } j\geq i\end{cases}$$

Notice that the model is non-parametric in that the parametric form of the survival distribution is not specified - preferably only the ordinal property of the survival times are used (in the determination of the risk sets). As this is a non-parametric case $\underset{\sim}{\varphi}$ is not required (ie $q=0$).

In another embodiment of the method of the invention, the model defined by the likelihood function is a Parametric survival model. Preferably, in a parametric survival model, $\underset{\sim}{\beta}^t$ is a vector of (explanatory) parameters associated with the components, and $\underset{\sim}{\varphi}^t$ is a vector of parameters associated with the functional form of the survival density function.

Preferably, the method of the invention provides for the parsimonious selection (and estimation) from the parameters $\underset{\sim}{\beta}^t$ and the estimation of $\underset{\sim}{\varphi}^t=(\varphi_1,\varphi_2,\cdots,\varphi_q)$ for parametric survival models given the data X, $\underset{\sim}{y}$ and $\underset{\sim}{c}$.

In applying a parametric survival model, the survival times do not require pre-processing and are denoted as $\underset{\sim}{y}$. The parametic survival model is applied as follows: Denote by $f(y;\underset{\sim}{\theta},\underset{\sim}{\beta},X)$ the parametric density function of the survival time, denote its survival function by

$\mathbb{S}(y;\underset{\sim}{\varphi},\underset{\sim}{\beta},X)=\int_y^\infty f(u;\underset{\sim}{\varphi},\underset{\sim}{\beta},X)du$ where $\underset{\sim}{\varphi}$ are the parameters relevant to the parametric form of the density function and $\underset{\sim}{\beta},X$ are as defined above. The hazard function is defined as $h(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X)=f(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X)/\mathbb{S}(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X)$.

Preferably, the generic formulation of the log-likelihood function, taking censored data into account, is

$$L=\sum_{i=1}^{N}\left\{c_i \log\left(f\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)\right)+(1-c_i)\log\left(\mathbb{S}\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)\right)\right\}$$

Reference to standard texts on analysis of survival time data via parametric regression survival models reveals a collection of survival time distributions that may be used. Survival distributions that may be used include, for example, the Weibull, Exponential or Extreme Value distributions.

If the hazard function may be written as

$h\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)=\lambda\left(y_i;\underset{\sim}{\varphi}\right)exp\left(X_i\underset{\sim}{\beta}\right)$ then $\mathbb{S}\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)=exp\left(-\Lambda\left(y_i;\underset{\sim}{\varphi}\right)e^{X_i\underset{\sim}{\beta}}\right)$ and

$f\left(y_i;\underset{\sim}{\varphi},\underset{\sim}{\beta},X\right)=\lambda\left(y_i;\underset{\sim}{\varphi}\right)exp\left(X_i\underset{\sim}{\beta}-\Lambda\left(y_i\right)e^{X_i\underset{\sim}{\beta}}\right)$ where $\Lambda\left(y_i;\underset{\sim}{\varphi}\right)=\int_{-\infty}^{y_i}\lambda\left(u;\underset{\sim}{\varphi}\right)du$

is the integrated hazard function and $\lambda\left(y_i;\underset{\sim}{\varphi}\right)=\frac{d\Lambda\left(y_i;\underset{\sim}{\varphi}\right)}{dy_i}$; X_i is the i^{th} row of X.

The Weibull, Exponential and Extreme Value distributions have density and hazard functions that may be written in the form of those presented in the paragraph immediately above.

The application detailed relies in part on an algorithm of Aitken and Clayton (1980) however it permits the user to specify any parametric underlying hazard function.

Following from Aitkin and Clayton (1980) a preferred likelihood function which models a parametic survival model is:

$$L=\sum_{i=1}^{N}\left\{c_i \log\left(\mu_i\right)-\mu_i+c_i\left(\log\left(\frac{\lambda\left(y_i\right)}{\Lambda\left(y_i;\underset{\sim}{\varphi}\right)}\right)\right)\right\} \qquad \text{(D5)}$$

where $\mu_i=\Lambda\left(y_i;\underset{\sim}{\varphi}\right)exp\left(X_i\underset{\sim}{\beta}\right)$. Aitkin and Clayton (1980) note that a consequence of equation (11) is that the c_i's may be

treatedas Poisson variates with means μ_i and that the last term in equation (11) does not depend on $\underset{\sim}{\beta}$ (although it depends on $\underset{\sim}{\varphi}$).

Preferably, the posterior distribution of $\underset{\sim}{\beta}$, $\underset{\sim}{\varphi}$ and $\underset{\sim}{\tau}$ given $\underset{\sim}{y}$ is

$$P(\underset{\sim}{\beta},\underset{\sim}{\varphi},\underset{\sim}{\tau}|\underset{\sim}{y}\underset{\sim}{c}) \propto l(\underset{\sim}{y}|\underset{\sim}{\beta},\underset{\sim}{\varphi},\underset{\sim}{c})P(\underset{\sim}{\beta}|\underset{\sim}{\tau})P(\underset{\sim}{\tau}) \qquad \text{(D6)}$$

wherein $l(\underset{\sim}{y}|\underset{\sim}{\beta},\underset{\sim}{\varphi},\underset{\sim}{c})$ is the likelihood function.

In one embodiment, $\underset{\sim}{\tau}$ may be treated as a vector of missing data and an iterative procedure used to maximise equation (D6) to produce a posteriori estimates of $\underset{\sim}{\beta}$. The prior of equation (D1) is such that the maximum a posteriori estimates will tend to be sparse i.e. if a large number of parameters are redundant, many components of $\underset{\sim}{\beta}$ will be zero.

Because a prior expectation exists that many components of $\underset{\sim}{\beta}^l$ are zero, the estimation may be performed in such a way that most of the estimated β_i's are zero and the remaining non-zero estimates provide an adequate explanation of the survival times.

In the context of microarray data this exercise translates to identifying a parsimonious set of genes that provide an adequate explanation for the event times.

As stated above, the component weights which maximise the posterior distribution may be determined using an iterative procedure. Preferable, the iterative procedure for maximising the posterior distribution of the components and component weights is an EM algorithm, such as, for example, that described in Dempster et al, 1977.

If the E step of the EM algorithm is examined, from (D6) ignoring terms not involving beta, it is necessary to compute

$$\sum_{i=1}^{n} E\{\beta_i^2/\tau_i^2 \mid \hat{\beta}_i\} = \sum_{i=1}^{n} \beta_i^2/\hat{d}_i^2 \qquad \text{(D7)}$$

where $\hat{d}_i = d_i(\hat{\beta}_i) = E\{1/\nu_i^2 \mid \hat{\beta}_i\}^{-0.5}$ and for convenience we define $\hat{d}_i = 1/\hat{d}_i = 0$ if $\hat{\beta}_i = 0$. In the following we write $\hat{d} = d(\hat{\beta}) = (\hat{d}_1, \hat{d}_2, ..., \hat{d}_n)^T$. In a similar manner we define for example $d(\beta^{(n)})$ and $d(\gamma^{(n)}) = P_n^T d(P_n \gamma^{(n)})$ where $\beta^{(n)} = P_n \gamma^{(n)}$ and P_n is obtained from the p by p identity matrix by omitting columns j for which $\beta_j^{(n)} = 0$.

Hence to do the E step requires the calculation of the conditional expected value of $t_i^2 = 1/\tau_i^2$ when $p\left(\beta_i \middle| \tau_i^2\right)$ is $N\left(0, \tau_i^2\right)$ and $p\left(\tau_i^2\right)$ has a specified prior distribution as discussed above.

In one embodiment, the EM algorithm comprises the steps:

1. Choose the hyperprior and values for its parameters, namely b and k. Initialising the algorithm by setting $n=0$, $S_0 = \{1, 2, ..., p\}$, initialise $\underset{\sim}{\beta}^{(0)} = \underset{\sim}{\beta}^*$, $\underset{\sim}{\varphi}^{(0)}$,

2. Defining

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

and let P_n be a matrix of zeroes and ones such that the nonzero elements $\underset{\sim}{\gamma}^{(n)}$ of $\underset{\sim}{\beta}^{(n)}$ satisfy

$$\underset{\sim}{\gamma}^{(n)} = P_n^T \underset{\sim}{\beta}^{(n)}, \quad \underset{\sim}{\beta}^{(n)} = P_n \underset{\sim}{\gamma}^{(n)}$$
$$\underset{\sim}{\gamma} = P_n^T \underset{\sim}{\beta}, \quad \underset{\sim}{\beta} = P_n \underset{\sim}{\gamma} \qquad \text{(D8)}$$

3. Performing an estimation step by calculating the expected value of the posterior distribution of component weights. This may be performed using the function:

$$Q(\underset{\sim}{\beta}\,|\,\underset{\sim}{\beta}^{(n)},\varphi^{(n)}) = E\left\{\log\left(P\left(\underset{\sim}{\beta},\underset{\sim}{\varphi},\tau\,|\,\underset{\sim}{y}\right)\right)|\,\underset{\sim}{y},\underset{\sim}{\beta}^{(n)},\varphi^{(n)}\right\}$$
$$= L(\underset{\sim}{y}\,|\,\underset{\sim}{\beta},\varphi^{(n)}) - \frac{1}{2}\sum_{i=1}^{p}\left(\frac{\beta_i}{d_i(\beta^{(n)})}\right)^2 \qquad \text{(D9)}$$

where L is the log likelihood function of $\underset{\sim}{y}$. Using $\beta = P_n\gamma$ and $\beta^{(n)} = P_n\gamma^{(n)}$ we have

$$Q(\underset{\sim}{\gamma}\,|\,\underset{\sim}{\gamma}^{(n)},\underset{\sim}{\varphi}^{(n)}) = L(\underset{\sim}{t}\,|\,P_n\underset{\sim}{\gamma},\underset{\sim}{\varphi}^{(n)}) - \frac{1}{2}\gamma^T\Delta(d(\gamma^{(n)}))\gamma \qquad \text{(D10)}$$

4. Performing the maximisation step. This may be performed using Newton Raphson iterations as follows: Set $\underset{\sim}{\gamma}_0 = \underset{\sim}{\gamma}^{(r)}$ and for r=0,1,2,... $\underset{\sim}{\gamma}_{r+1} = \underset{\sim}{\gamma}_r + \alpha_r\,\underset{\sim}{\delta}_r$ where α_r is chosen by a line search algorithm to ensure $Q\left(\underset{\sim}{\gamma}_{r+1}\,|\,\underset{\sim}{\gamma}^{(n)},\underset{\sim}{\varphi}^{(n)}\right) > Q\left(\underset{\sim}{\gamma}_r\,|\,\underset{\sim}{\gamma}^{(n)},\underset{\sim}{\varphi}^{(n)}\right)$, and

$$\underset{\sim}{\delta}_r = \Delta(d(\underset{\sim}{\gamma}^{(n)}))[-\Delta(d(\underset{\sim}{\gamma}^{(n)}))\frac{\partial^2 L}{\partial^2\underset{\sim}{\gamma}_r}\Delta(d(\underset{\sim}{\gamma}^{(n)})) + I]^{-1}(\frac{\partial L}{\partial\underset{\sim}{\gamma}_r} - \frac{\underset{\sim}{\gamma}_r}{d(\underset{\sim}{\gamma}^{(n)})})$$

where $\frac{\partial L}{\partial\underset{\sim}{\gamma}_r} = P_n^T\frac{\partial L}{\partial\underset{\sim}{\beta}_r}$, $\frac{\partial^2 L}{\partial^2\underset{\sim}{\gamma}_r} = P_n^T\frac{\partial^2 L}{\partial^2\underset{\sim}{\beta}_r}P_n$ for $\underset{\sim}{\beta}_r = P_n\underset{\sim}{\gamma}_r$ (D11)

Let $\underset{\sim}{\gamma}^*$ be the value of $\underset{\sim}{\gamma}_r$ when some convergence criterion is satisfied e.g $||\underset{\sim}{\gamma}_r - \underset{\sim}{\gamma}_{r+1}|| < \varepsilon$ (for example $\varepsilon = 10^{-5}$)

5. Define $\underset{\sim}{\beta}^* = P_n\underset{\sim}{\gamma}^*$, $S_n = \left\{i : |\beta_i| > \varepsilon_1 \max_j|\beta_j|\right\}$ where ε_1 is a small constant, say 10^{-5}. Set $n=n+1$, choose $\underset{\sim}{\varphi}^{(n+1)} = \underset{\sim}{\varphi}^{(n)} + \kappa_n\left(\underset{\sim}{\varphi}^* - \underset{\sim}{\varphi}^{(n)}\right)$

where $\underset{\sim}{\varphi}^*$ satisfies $\frac{\partial L\left(\underset{\sim}{y} \mid P_n \underset{\sim}{\gamma}^*, \underset{\sim}{\varphi}\right)}{\partial \underset{\sim}{\varphi}} = 0$ and κ_n is a damping factor such that $0 < \kappa_n < 1$.

6. Check convergence. If $\|\underset{\sim}{\gamma}^* - \underset{\sim}{\gamma}^{(n)}\| < \varepsilon_2$ where ε_2 is suitably small then stop, else go to step 2 above.

In another embodiment, step (D11) in the maximisation step may be estimated by replacing $\frac{\partial^2 L}{\partial^2 \gamma_r}$ with its expectation $E\{\frac{\partial^2 L}{\partial^2 \gamma_r}\}$.

In one embodiment, the EM algorithm is applied to maximise the posterior distribution when the model is Cox's proportional hazard's model.

To aid in the exposition of the application of the EM algorithm when the model is Cox's proportional hazards model, it is preferred to define "dynamic weights" and matrices based on these weights. The weights are -

$$w_{i,l} = \frac{\zeta_{i,l} \exp\left(Z_l \underset{\sim}{\beta}\right)}{\sum_{j=1}^{N} \zeta_{i,j} \exp\left(Z_j \underset{\sim}{\beta}\right)},$$

$$w_l^* = \sum_{i=1}^{N} d_i w_{i,l},$$

$$\tilde{w}_l = d_l - w_l^*.$$

Matrices based on these weights are -

$$W_i = \begin{pmatrix} w_{i,1} \\ w_{i,2} \\ \vdots \\ \vdots \\ w_{i,N} \end{pmatrix},$$

$$\tilde{W} = \begin{pmatrix} \tilde{w}_1 \\ \tilde{w}_2 \\ \vdots \\ \vdots \\ \tilde{w}_N \end{pmatrix},$$

$$\Delta\left(W^*\right) = \begin{pmatrix} w_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_N^* \end{pmatrix},$$

$$W^{**} = \sum_{i=1}^{N} d_i W_i W_i^T$$

In terms of the matrices of weights the first and second derivatives of L may be written as

$$\left.\begin{aligned} \frac{\partial L}{\partial \underset{\sim}{\beta}} &= Z^T \tilde{W} \\ \frac{\partial^2 L}{\partial \underset{\sim}{\beta}^2} &= Z^T \left(W^{**} - \Delta\left(W^*\right)\right) Z = Z^T K Z \end{aligned}\right\} \qquad \text{(D12)}$$

where $K = W - \Delta(W)$. Note therefore from the transformation matrix P_n described as part of Step (2) of the EM algorithm (Equation D8) (see also Equations (D11)) it follows that

$$\left.\begin{aligned} \frac{\partial L}{\partial \underset{\sim}{\gamma}_r} &= P_n^T \frac{\partial L}{\partial \underset{\sim}{\beta}_r} = P_n^T Z^T \tilde{W} \\ \frac{\partial^2 L}{\partial \underset{\sim}{\gamma}_r^2} &= P_n^T \frac{\partial^2 L}{\partial \underset{\sim}{\beta}_r^2} P_n = P_n^T Z^T \left(W^{**} - \Delta\left(W^*\right)\right) Z P_n = P_n^T Z^T K Z P_n \end{aligned}\right\} \qquad \text{(D13)}$$

Preferably, when the model is Cox's proportional hazards model the E step and M step of the EM algorithm are as follows:

1. Choose the hyperprior and its parameters b and k. Set $n=0$, $S_0 = \{1,2,\cdots, p\}$. Let v be the vector with components

$$v_i = \begin{cases} 1-\epsilon, & \text{if } c_i=1 \\ \epsilon, & \text{if } c_i=0 \end{cases}$$

for some small ε, say .001. Define f to be log(v/t).

If $p \leq N$ compute initial values $\underset{\sim}{\beta}^*$ by

$$\underset{\sim}{\beta}^* = (Z^T Z + \lambda I)^{-1} Z^T \mathrm{f} \quad .$$

If $p > N$ compute initial values $\underset{\sim}{\beta}^*$ by

$$\underset{\sim}{\beta}^* = \frac{1}{\lambda}(I - Z^T(ZZ^T + \lambda I)^{-1} Z) Z^T \mathrm{f}$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$.

2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

Let P_n be a matrix of zeroes and ones such that the nonzero elements $\underset{\sim}{\gamma}^{(n)}$ of $\underset{\sim}{\beta}^{(n)}$ satisfy

$$\underset{\sim}{\gamma}^{(n)} = P_n^T \underset{\sim}{\beta}^{(n)} \quad , \quad \underset{\sim}{\beta}^{(n)} = P_n \underset{\sim}{\gamma}^{(n)}$$
$$\underset{\sim}{\gamma} = P_n^T \underset{\sim}{\beta} \quad , \quad \underset{\sim}{\beta} = P_n \underset{\sim}{\gamma}$$

3. Perform the E step by calculating

$$Q(\underset{\sim}{\beta}\,|\,\underset{\sim}{\beta}^{(n)}) = E\left\{\log\left(P\left(\underset{\sim}{\beta},\varphi,\tau\,|\,\underset{\sim}{t}\right)\right)|\,\underset{\sim}{t},\underset{\sim}{\beta}^{(n)}\right\}$$

$$= L(\underset{\sim}{t}\,|\,\underset{\sim}{\beta}) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\beta_i}{d_i(\beta^{(n)})}\right)^2$$

where L is the log likelihood function of $\underset{\sim}{t}$ given by Equation (8). Using $\beta = P_n\gamma$ and $\beta^{(n)} = P_n\gamma^{(n)}$ we have

$$Q(\underset{\sim}{\gamma}\,|\,\underset{\sim}{\gamma}^{(n)}) = L(\underset{\sim}{t}\,|\,P_n\underset{\sim}{\gamma}) - \frac{1}{2}\gamma^T\Delta(d(\gamma^{(n)}))\gamma$$

4. Do the M step. This can be done with Newton Raphson iterations as follows. Set $\underset{\sim}{\gamma}_0 = \underset{\sim}{\gamma}^{(r)}$ and for r=0,1,2,... $\underset{\sim}{\gamma}_{r+1} = \underset{\sim}{\gamma}_r + \alpha_r\,\underset{\sim}{\delta}_r$ where α_r is chosen by a line search algorithm to ensure $Q\left(\underset{\sim}{\gamma}_{r+1}\,|\,\underset{\sim}{\gamma}^{(n)},\underset{\sim}{\varphi}^{(n)}\right) > Q\left(\underset{\sim}{\gamma}_r\,|\,\underset{\sim}{\gamma}^{(n)},\underset{\sim}{\varphi}^{(n)}\right)$.

For $p \leq N$ use

$$\underset{\sim}{\delta}_r = \Delta\left(d(\underset{\sim}{\gamma}^{(n)})\right)\left(Y^TKY+I\right)^{-1}\left(Y^T\tilde{W} - \Delta\left(1/d(\underset{\sim}{\gamma}^{(n)})\right)\underset{\sim}{\gamma}\right),$$

where $Y = ZP_n\Delta\left(d(\underset{\sim}{\gamma}^{(n)})\right)$.

For p > N use

$$\underset{\sim}{\delta}_r = \Delta\left(d(\underset{\sim}{\gamma}^{(n)})\right)\left(I - Y^T\left(YY^T+K^{-1}\right)^{-1}Y\right)\left(Y^T\tilde{W} - \Delta\left(1/d(\underset{\sim}{\gamma}^{(n)})\right)\underset{\sim}{\gamma}\right)$$

Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g $||\,\gamma_r - \gamma_{r+1}\,|| < \varepsilon$ (for example 10^{-5}).

5. Define $\underset{\sim}{\beta}^* = P_n\underset{\sim}{\gamma}^*$, $S_n = \left\{i : |\,\beta_i\,| > \varepsilon_1 \max_j |\beta_j|\right\}$ where ε_1 is a small constant, say 10^{-5}. This step eliminates variables with very small coefficients.

6. Check convergence. If $||\underset{\sim}{\gamma}^* - \underset{\sim}{\gamma}^{(n)}|| < \varepsilon_2$ where ε_2 is suitably small then stop, else set n=n+1, go to step 2 above and repeat procedure until convergence occurs.

In another embodiment the EM algorithm is applied to maximise the posterior distribution when the model is a parametric survival model.

In applying the EM algorithm to the parametic survival model, a consequence of equation (11) is that the c_i's may be treated as Poisson variates with means μ_{i1} and that the last term in equation (11) does not depend on β (although it depends on φ). Note that $log(\mu_i) = log\left(\Lambda\left(y_i;\underset{\sim}{\varphi}\right)\right) + X_i\underset{\sim}{\beta}$ and so it is possible to couch the problem in terms of a log-linear model for the Poisson-like mean. Preferably, an iterative maximization of the log-likelihood function is performed where given initial estimates of $\underset{\sim}{\varphi}$ the estimates of $\underset{\sim}{\beta}$ are obtained. Then given these estimates of $\underset{\sim}{\beta}$, updated estimates of $\underset{\sim}{\varphi}$ are obtained. The procedure is continued until convergence occurs.

Applying the posterior distribution described above, we note that (for fixed $\underset{\sim}{\varphi}$)

$$\frac{\partial log(\mu)}{\partial \underset{\sim}{\beta}} = \frac{1}{\mu}\frac{\partial \mu}{\partial \underset{\sim}{\beta}} \Leftrightarrow \frac{\partial \mu}{\partial \underset{\sim}{\beta}} = \mu\frac{\partial log(\mu)}{\partial \underset{\sim}{\beta}} \text{ and } \frac{\partial log(\mu_i)}{\partial \underset{\sim}{\beta}} = X_i \qquad \text{(D14)}$$

Consequently from Equations (11) and (12) it follows that $\frac{\partial L}{\partial \underset{\sim}{\beta}} = X^T\left(\underset{\sim}{c} - \underset{\sim}{\mu}\right)$ and $\frac{\partial^2 L}{\partial \underset{\sim}{\beta}^2} = -X^T\Delta\left(\underset{\sim}{\mu}\right)X$.

The versions of Equation (12) relevant to the parametric survival models are

$$\left.\begin{aligned} \frac{\partial L}{\partial \underset{\sim}{\gamma}_r} &= P_n^T\frac{\partial L}{\partial \underset{\sim}{\beta}_r} = P_n^T X^T\left(\underset{\sim}{c} - \underset{\sim}{\mu}\right) \\ \frac{\partial^2 L}{\partial \underset{\sim}{\gamma}_r^2} &= P_n^T\frac{\partial^2 L}{\partial \underset{\sim}{\beta}_r^2}P_n = -P_n^T X^T\Delta\left(\underset{\sim}{\mu}\right)XP_n \end{aligned}\right\} \qquad \text{(D15)}$$

To solve for $\underset{\sim}{\varphi}$ after each M step of the EM algorithm (see step 5 below) preferably put $\underset{\sim}{\varphi}^{(n+1)} = \underset{\sim}{\varphi}^{(n+1)} + \kappa_n\left(\underset{\sim}{\varphi}^* - \underset{\sim}{\varphi}^{(n)}\right)$ where $\underset{\sim}{\varphi}^*$

satisfies $\frac{\partial L}{\partial \underset{\sim}{\varphi}}=0$ for $0<\kappa_n \leq 1$ and β is fixed at the value obtained from the previous M step.

It is possible to provide an EM algorithm for parameter selection in the context of parametric survival models and microarray data. Preferably, the EM algorithm is as follows:

1. Choose a hyper prior and its parameters b and k eg b=1e7 and k=0.5. Set $n=0$, $S_0 = \{1,2,\cdots, p\}$ $\underset{\sim}{\varphi}^{(initial)} = \underset{\sim}{\varphi}^{(0)}$. Let v be the vector with components

$$v_i = \begin{cases} 1-\epsilon, & if \quad c_i=1 \\ \epsilon, & if \quad c_i=0 \end{cases}$$

for some small ε, say for example .001. Define f to be $\log(v/\Lambda(y,\varphi))$.

If $p \leq N$ compute initial values $\underset{\sim}{\beta}^*$ by $\underset{\sim}{\beta}^* = (X^T X + \lambda I)^{-1} X^T \mathrm{f}$.
If $p > N$ compute initial values $\underset{\sim}{\beta}^*$ by

$$\underset{\sim}{\beta}^* = \frac{1}{\lambda}(I - X^T(XX^T + \lambda I)^{-1}X)X^T \mathrm{f}$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$.

2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0, & \text{otherwise} \end{cases}$$

Let P_n be a matrix of zeroes and ones such that the nonzero elements $\underset{\sim}{\gamma}^{(n)}$ of $\underset{\sim}{\beta}^{(n)}$ satisfy

$$\underset{\sim}{\gamma}^{(n)} = P_n^T \underset{\sim}{\beta}^{(n)}, \quad \underset{\sim}{\beta}^{(n)} = P_n \underset{\sim}{\gamma}^{(n)}$$
$$\underset{\sim}{\gamma} = P_n^T \underset{\sim}{\beta}, \quad \underset{\sim}{\beta} = P_n \underset{\sim}{\gamma}$$

3. Perform the E step by calculating

$$Q(\underset{\sim}{\beta}\,|\,\underset{\sim}{\beta}^{(n)},\underset{\sim}{\varphi}^{(n)}) = E\left\{\log\left(P\left(\underset{\sim}{\beta},\underset{\sim}{\varphi},\tau\,|\,\underset{\sim}{y}\right)\right)|\,\underset{\sim}{y},\underset{\sim}{\beta}^{(n)},\underset{\sim}{\varphi}^{(n)}\right\}$$

$$= L(\underset{\sim}{y}\,|\,\underset{\sim}{\beta},\underset{\sim}{\varphi}^{(n)}) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\beta_i}{d(\beta^{(n)})}\right)^2$$

where L is the log likelihood function of $\underset{\sim}{y}$ and $\underset{\sim}{\varphi}^{(n)}$.

Using $\beta = P_n\gamma$ and $\beta^{(n)} = P_n\gamma^{(n)}$ we have

$$Q(\underset{\sim}{\gamma}\,|\,\underset{\sim}{\gamma}^{(n)},\underset{\sim}{\varphi}^{(n)}) = L(\underset{\sim}{y}\,|\,P_n\underset{\sim}{\gamma},\underset{\sim}{\varphi}^{(n)}) - \frac{1}{2}\gamma^T\Delta(d(\gamma^{(n)}))\gamma$$

4. Do the M step. This can be done with Newton Raphson iterations as follows. Set $\underset{\sim}{\gamma}_0 = \underset{\sim}{\gamma}^{(r)}$ and for r=0,1,2,...

$\underset{\sim}{\gamma}_{r+1} = \underset{\sim}{\gamma}_r + \alpha_r\,\underset{\sim}{\delta}_r$ where α_r is chosen by a line search algorithm to ensure $Q\left(\underset{\sim}{\gamma}_{r+1}\,|\,\underset{\sim}{\gamma}^{(n)},\underset{\sim}{\varphi}^{(n)}\right) > Q\left(\underset{\sim}{\gamma}_r\,|\,\underset{\sim}{\gamma}^{(n)},\underset{\sim}{\varphi}^{(n)}\right)$.

For $p \leq N$ use

$$\underset{\sim}{\delta}_r = -\Delta(d(\underset{\sim}{\gamma}^{(n)}))[Y_n^T\Delta\left(\underset{\sim}{\mu}\right)Y_n + I]^{-1}(Y_n^T\left(\underset{\sim}{c}-\underset{\sim}{\mu}\right) - \Delta\left(1/d(\underset{\sim}{\gamma}^{(n)})\right)\underset{\sim}{\gamma})$$

where $Y = XP_n\Delta\left(d(\underset{\sim}{\gamma}^{(n)})\right)$.

For p > N use

$$\underset{\sim}{\delta}_r = -\Delta\left(d(\underset{\sim}{\gamma}^{(n)})\right)\left(I - Y^T\left(YY^T + \Delta\left(1/\underset{\sim}{\mu}\right)\right)^{-1}Y\right)\left(Y_n^T\left(\underset{\sim}{c}-\underset{\sim}{\mu}\right) - \Delta\left(1/d(\underset{\sim}{\gamma}^{(n)})\right)\underset{\sim}{\gamma}\right)$$

Let γ^* be the value of γ_r when some convergence criterion is satisfied e.g $||\ \gamma_r - \gamma_{r+1}\ ||\ < \varepsilon$ (for example 10^{-5}).

5. Define $\underset{\sim}{\beta}^* = P_n\underset{\sim}{\gamma}^*$, $S_n = \left\{i : |\,\beta_i\,| > \varepsilon_1 \max_j |\beta_j|\right\}$ where ε_1 is a small constant, say 10^{-5}. Set n=n+1, choose $\underset{\sim}{\varphi}^{(n+1)} = \underset{\sim}{\varphi}^{(n)} + \kappa_n\left(\underset{\sim}{\varphi}^* - \underset{\sim}{\varphi}^{(n)}\right)$

where $\underset{\sim}{\varphi}^*$ satisfies $\frac{\partial L\left(\underset{\sim}{y}|P_n\underset{\sim}{\gamma}^*,\underset{\sim}{\varphi}\right)}{\partial\underset{\sim}{\varphi}}=0$ and κ_n is a damping factor such that $0<\kappa_n<1$.

6. Check convergence. If $||\underset{\sim}{\gamma}^{\tau}-\underset{\sim}{\gamma}^{(n)}||<\varepsilon_2$ where ε_2 is suitably small then stop, else go to step 2.

In another embodiment, survival times are described by a Weibull survival density function. For the Weibull case $\underset{\sim}{\varphi}$ is preferably one dimensional and

$$\Lambda\left(y;\underset{\sim}{\varphi}\right)=y^{\alpha},$$
$$\lambda\left(y;\underset{\sim}{\varphi}\right)=\alpha y^{\alpha-1},$$
$$\underset{\sim}{\varphi}=\alpha$$

Preferably, $\frac{\partial L}{\partial\alpha}=\frac{N}{\alpha}+\sum_{i=1}^{N}(c_i-\mu_i)log(y_i)=0$ is solved after each M step so as to provide an updated value of α.

Following the steps applied for Cox's proportional hazards model, one may estimate α and select a parsimonious subset of parameters from $\underset{\sim}{\beta}$ that can provide an adequate explanation for the survival times if the survival times follow a Weibull distribution. A numerical example is now given.

The preferred embodiment of the present invention will now be described by way of reference only to the following non-limiting example. It should be understood, however, that the example is illustrative only, should not be taken in any way as a restriction on the generality of the invention described above.

Example:

Full normal regression example 201 data points 41 basis functions

k=0 and b=1e7
the correct four basis functions are identified namely
2 12 24 34
estimated variance is 0.67.

With k=0.2 and b=1e7
eight basis functions are identified, namely
2 8 12 16 19 24 34
estimated variance is 0.63. Note that the correct set of basis functions is included in this set.

The results of the iterations for k=0.2 and b=1e7 are given below.

EM Iteration: 0 expected post: 2 basis fns 41

sigma squared 0.6004567
EM Iteration: 1 expected post: -63.91024 basis fns 41

sigma squared 0.6037467
EM Iteration: 2 expected post: -52.76575 basis fns 41

sigma squared 0.6081233
EM Iteration: 3 expected post: -53.10084 basis fns 30

sigma squared 0.6118665
EM Iteration: 4 expected post: -53.55141 basis fns 22

sigma squared 0.6143482
EM Iteration: 5 expected post: -53.79887 basis fns 18

sigma squared 0.6155
EM Iteration: 6 expected post: -53.91096 basis fns 18

```
sigma squared 0.6159484
EM Iteration:  7 expected post:  -53.94735 basis fns 16

sigma squared 0.6160951
EM Iteration:  8 expected post:  -53.92469 basis fns 14

sigma squared 0.615873
EM Iteration:  9 expected post:  -53.83668 basis fns 13

sigma squared 0.6156233
EM Iteration:  10 expected post:  -53.71836 basis fns 13

sigma squared 0.6156616
EM Iteration:  11 expected post:  -53.61035 basis fns 12

sigma squared 0.6157966
EM Iteration:  12 expected post:  -53.52386 basis fns 12

sigma squared 0.6159524
EM Iteration:  13 expected post:  -53.47354 basis fns 12

sigma squared 0.6163736
EM Iteration:  14 expected post:  -53.47986 basis fns 12

sigma squared 0.6171314
EM Iteration:  15 expected post:  -53.53784 basis fns 11

sigma squared 0.6182353
EM Iteration:  16 expected post:  -53.63423 basis fns 11

sigma squared 0.6196385
EM Iteration:  17 expected post:  -53.75112 basis fns 11

sigma squared 0.621111
EM Iteration:  18 expected post:  -53.86309 basis fns 11

sigma squared 0.6224584
```

```
EM Iteration:  19 expected post:  -53.96314 basis fns 11

sigma squared 0.6236203
EM Iteration:  20 expected post:  -54.05662 basis fns 11

sigma squared 0.6245656
EM Iteration:  21 expected post:  -54.1382 basis fns 10

sigma squared 0.6254182
EM Iteration:  22 expected post:  -54.21169 basis fns 10

sigma squared 0.6259266
EM Iteration:  23 expected post:  -54.25395 basis fns 9

sigma squared 0.6259266
EM Iteration:  24 expected post:  -54.26136 basis fns 9

sigma squared 0.6260238
EM Iteration:  25 expected post:  -54.25962 basis fns 9

sigma squared 0.6260203
EM Iteration:  26 expected post:  -54.25875 basis fns 8

sigma squared 0.6260179
EM Iteration:  27 expected post:  -54.25836 basis fns 8

sigma squared 0.626017
EM Iteration:  28 expected post:  -54.2582 basis fns 8

sigma squared 0.6260166
```

For the reduced data set with 201 observations and 10 variables, k=0 and b=1e7
Gives the correct basis functions, namely 1 2 3 4. With k=0.25 and b=1e7, 7 basis functions were chosen, namely 1 2 3 4 6 8 9. A record of the iterations is given below. Note that this set also includes the correct set.

EM Iteration: 0 expected post: 2 basis fns 10

sigma squared 0.6511711
EM Iteration: 1 expected post: -66.18153 basis fns 10

sigma squared 0.6516289
EM Iteration: 2 expected post: -57.69118 basis fns 10

sigma squared 0.6518373
EM Iteration: 3 expected post: -57.72295 basis fns 9

sigma squared 0.6518373
EM Iteration: 4 expected post: -57.74616 basis fns 8

sigma squared 0.65188
EM Iteration: 5 expected post: -57.75293 basis fns 7

sigma squared 0.6518781

Ordered categories examples

Luo prostate data 15 samples 9605 genes. For k=0 and b=1e7 we get the following results

misclassification table

	pred			
y	1	2	3	4
1	4	0	0	0
2	0	2	1	0
3	0	0	4	0
4	0	0	0	4

Identifiers of variables left in ordered categories model
6611

For k=0.25 and b=1e7 we get the following results

misclassification table

```
      pred
   y   1 2 3 4
     1 4 0 0 0
     2 0 3 0 0
     3 0 0 4 0
     4 0 0 0 4
```

Identifiers of variables left in ordered categories model
6611 7188

Note that we now have perfect discrimination on the training data with the addition of one extra variable. A record of the iterations of the algorithm is given below.

```
**************************************************
Iteration 1  :  11 cycles, criterion -4.661957

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class 1 Number of basis functions in model : 9608
**************************************************
Iteration 2  :  5 cycles, criterion -9.536942

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  1 21
row =true class

Class 1 Number of basis functions in model : 6431
**************************************************
```

```
Iteration 3   :  4 cycles, criterion -9.007843

misclassification matrix
    fhat
f     1  2
  1 23  0
  2  0 22
row =true class

Class 1 Number of basis functions in model : 508
**************************************************
Iteration 4   :  5 cycles, criterion -6.47555

misclassification matrix
    fhat
f     1  2
  1 23  0
  2  0 22
row =true class

Class 1 Number of basis functions in model : 62
**************************************************
Iteration 5   :  6 cycles, criterion -4.126996

misclassification matrix
    fhat
f     1  2
  1 23  0
  2  1 21
row =true class

Class 1 Number of basis functions in model : 20
**************************************************
Iteration 6   :  6 cycles, criterion -3.047699

misclassification matrix
    fhat
```

```
f     1  2
  1 23  0
  2  1 21
row =true class

Class 1 Number of basis functions in model : 12
************************************************
Iteration 7  :  5 cycles, criterion -2.610974

misclassification matrix
   fhat
f     1  2
  1 23  0
  2  1 21
row =true class

Class  1 : Variables left in model
1 2 3 408 846 6614 7191 8077
regression coefficients
28.81413 14.27784 7.025863 -1.086501e-06 4.553004e-09 -
16.25844 0.1412991 -0.04101412

************************************************
Iteration 8  :  5 cycles, criterion -2.307441

misclassification matrix
   fhat
f     1  2
  1 23  0
  2  1 21
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191 8077
regression coefficients
32.66699 15.80614 7.86011 -18.53527 0.1808061 -0.006728619
```

```
**************************************************
Iteration 9   :  5 cycles, criterion -2.028043

misclassification matrix
    fhat
f     1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191 8077
regression coefficients
36.11990 17.21351 8.599812 -20.52450 0.2232955 -0.0001630341

**************************************************
Iteration 10   :  6 cycles, criterion -1.808861

misclassification matrix
    fhat
f     1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191 8077
regression coefficients
39.29053 18.55341 9.292612 -22.33653 0.260273 -8.696388e-08

**************************************************
Iteration 11   :  6 cycles, criterion -1.656129

misclassification matrix
    fhat
f     1  2
  1 23  0
```

```
  2   0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
42.01569 19.73626 9.90312 -23.89147 0.2882204

**************************************************
Iteration 12  :  6 cycles, criterion -1.554494

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
44.19405 20.69926 10.40117 -25.1328 0.3077712
**************************************************
Iteration 13  :  6 cycles, criterion -1.487778

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
45.84032 21.43537 10.78268 -26.07003 0.3209974
```

```
**************************************************
Iteration 14  :  6 cycles, criterion -1.443949

misclassification matrix
     fhat
f     1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
47.03702 21.97416 11.06231 -26.75088 0.3298526

**************************************************
Iteration 15  :  6 cycles, criterion -1.415

misclassification matrix
     fhat
f     1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
47.88472 22.35743 11.26136 -27.23297 0.3357765

**************************************************
Iteration 16  :  6 cycles, criterion -1.395770

misclassification matrix
     fhat
f     1  2
  1 23  0
```

```
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
48.47516 22.62508 11.40040 -27.56866 0.3397475

************************************************
Iteration 17  :  5 cycles, criterion -1.382936

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
48.88196 22.80978 11.49636 -27.79991 0.3424153

************************************************
Iteration 18  :  5 cycles, criterion -1.374340

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.16029 22.93629 11.56209 -27.95811 0.3442109
```

```
**************************************************
Iteration 19  :  5 cycles, criterion -1.368567

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.34987 23.02251 11.60689 -28.06586 0.3454208

**************************************************
Iteration 20  :  5 cycles, criterion -1.364684

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.47861 23.08109 11.63732 -28.13903 0.3462368

**************************************************
Iteration 21  :  5 cycles, criterion -1.362068

misclassification matrix
   fhat
f    1  2
```

```
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.56588 23.12080 11.65796 -28.18862 0.3467873

**************************************************
Iteration 22  :  5 cycles, criterion -1.360305

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.62496 23.14769 11.67193 -28.22219 0.3471588

**************************************************
Iteration 23  :  4 cycles, criterion -1.359116

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
```

```
49.6649 23.16588 11.68137 -28.2449 0.3474096

**************************************************
Iteration 24  :  4 cycles, criterion -1.358314

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.69192 23.17818 11.68776 -28.26025 0.3475789

**************************************************
Iteration 25  :  4 cycles, criterion -1.357772

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.71017 23.18649 11.69208 -28.27062 0.3476932

**************************************************
Iteration 26  :  4 cycles, criterion -1.357407

misclassification matrix
   fhat
```

```
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.72251 23.19211 11.695 -28.27763 0.3477704

************************************************
Iteration 27  :  4 cycles, criterion -1.35716

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.73084 23.19590 11.69697 -28.28237 0.3478225

************************************************
Iteration 28  :  3 cycles, criterion -1.356993

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
```

```
regression coefficients
49.73646 23.19846 11.6983 -28.28556 0.3478577

**************************************************
Iteration 29  :  3 cycles, criterion -1.356881

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.74026 23.20019 11.6992 -28.28772 0.3478814

**************************************************
Iteration 30  :  3 cycles, criterion -1.356805

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 6614 7191
regression coefficients
49.74283 23.20136 11.69981 -28.28918 0.3478975

1

misclassification table
   pred
```

y	1	2	3	4
1	4	0	0	0
2	0	3	0	0
3	0	0	4	0
4	0	0	0	4

Identifiers of variables left in ordered categories model
6611 7188

Ordered categories example
Luo prostate data 15 samples 50 genes. For k=0 and b=1e7 we get the following results

misclassification table

	pred			
y	1	2	3	4
1	4	0	0	0
2	0	2	1	0
3	0	0	4	0
4	0	0	0	4

Identifiers of variables left in ordered categories model
1

For k=0.25 and b=1e7 we get the following results

misclassification table

	pred			
y	1	2	3	4
1	4	0	0	0
2	0	3	0	0
3	0	0	4	0
4	0	0	0	4

Identifiers of variables left in ordered categories model
1 42

A record of the iterations for k=0.25 and b=1e7 is given below

```
**************************************************
Iteration 1  :  19 cycles, criterion -0.4708706

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class 1 Number of basis functions in model : 53
**************************************************
Iteration 2  :  7 cycles, criterion -1.536822

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class 1 Number of basis functions in model : 53
**************************************************
Iteration 3  :  5 cycles, criterion -2.032919

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class 1 Number of basis functions in model : 42
**************************************************
```

```
Iteration 4  :  5 cycles, criterion -2.132546

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class 1 Number of basis functions in model : 18
************************************************
Iteration 5  :  5 cycles, criterion -1.978462

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class 1 Number of basis functions in model : 13
************************************************
Iteration 6  :  5 cycles, criterion -1.668212

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 10 41 43 45
regression coefficients
34.13253 22.30781 13.04342 -16.23506 0.003213167 0.006582334
-0.0005943874 -3.557023
```

```
**************************************************
Iteration 7  :  5 cycles, criterion -1.407871

misclassification matrix
    fhat
f     1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 10 41 43 45
regression coefficients
36.90726 24.69518 14.61792 -17.16723 1.112172e-05 5.949931e-
06 -3.892181e-08 -4.2906

**************************************************
Iteration 8  :  5 cycles, criterion -1.244166

misclassification matrix
    fhat
f     1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 10 45
regression coefficients
39.15038 26.51011 15.78594 -17.99800 1.125451e-10 -4.799167

**************************************************
Iteration 9  :  5 cycles, criterion -1.147754

misclassification matrix
    fhat
f     1  2
```

```
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
40.72797 27.73318 16.56101 -18.61816 -5.115492

**************************************************
Iteration 10  :  5 cycles, criterion -1.09211

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
41.74539 28.49967 17.04204 -19.03293 -5.302421

**************************************************
Iteration 11  :  5 cycles, criterion -1.060238

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
```

```
42.36866 28.96076 17.32967 -19.29261 -5.410496

************************************************
Iteration 12  :  5 cycles, criterion -1.042037

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
42.73908 29.23176 17.49811 -19.44894 -5.472426

************************************************
Iteration 13  :  5 cycles, criterion -1.031656

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
42.95536 29.38894 17.59560 -19.54090 -5.507787

************************************************
Iteration 14  :  4 cycles, criterion -1.025738

misclassification matrix
   fhat
```

```
f     1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
43.08034 29.47941 17.65163 -19.59428 -5.527948

************************************************
Iteration 15  :  4 cycles, criterion -1.022366

misclassification matrix
    fhat
f     1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
43.15213 29.53125 17.68372 -19.62502 -5.539438

************************************************
Iteration 16  :  4 cycles, criterion -1.020444

misclassification matrix
    fhat
f     1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
```

```
regression coefficients
43.19322 29.56089 17.70206 -19.64265 -5.545984

************************************************
Iteration 17  :  4 cycles, criterion -1.019349

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
43.21670 29.57780 17.71252 -19.65272 -5.549713

************************************************
Iteration 18  :  3 cycles, criterion -1.018725

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
43.23008 29.58745 17.71848 -19.65847 -5.551837

************************************************
Iteration 19  :  3 cycles, criterion -1.01837

misclassification matrix
```

```
      fhat
     f    1  2
       1 23  0
       2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
43.23772 29.59295 17.72188 -19.66176 -5.553047

************************************************
Iteration 20  :  3 cycles, criterion -1.018167

misclassification matrix
      fhat
     f    1  2
       1 23  0
       2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
43.24208 29.59608 17.72382 -19.66363 -5.553737

************************************************
Iteration 21  :  3 cycles, criterion -1.018052

misclassification matrix
      fhat
     f    1  2
       1 23  0
       2  0 22
row =true class

Class  1 : Variables left in model
```

```
1 2 3 4 45
regression coefficients
43.24456 29.59787 17.72493 -19.66469 -5.55413

**************************************************
Iteration 22  :  3 cycles, criterion -1.017986

misclassification matrix
   fhat
f    1  2
  1 23  0
  2  0 22
row =true class

Class  1 : Variables left in model
1 2 3 4 45
regression coefficients
43.24598 29.59889 17.72556 -19.6653 -5.554354

1

misclassification table
   pred
y   1 2 3 4
  1 4 0 0 0
  2 0 3 0 0
  3 0 0 4 0
  4 0 0 0 4
Identifiers of variables left in ordered categories model
1 42
```